福 井 大 学 審 査

学位論文 [博士 (工学)]

A Dissertation Submitted to the University of Fukui for the Degree of Doctor of Engineering

Research on Interactive Teaching System with Eye Tracking Interface Based on Intention Recognition of Human

（視線追跡による人間の意図認識に基づく
インタラクティブ教示システムに関する研究)

2014年9月

王　茂
Mao Wang

Abstract

HCI (human-computer interaction) has been a major area of research in computer science, human factors, engineering psychology and closely related disciplines. And current interaction methods tend to be one side, with the bandwidth from the computer to the user far greater than that from user to computer. A fast and effortless mode of communication from a user to a computer would help redress this imbalance. Recently, eye tracking technology, with its special advantages of more direct interaction and freeing up the hands, has been widely studied. At the same time, because it can be applied to such as ALS (amyotrophic lateral sclerosis) patients, it is also useful as a solution for disabilities or patients to interact with computer and others.

The technology for measuring eye movements by computing pupil position in real time has been improving. The problem still exist is how to find an appropriate interaction method that can reflect user's intention or the main purpose of interaction. Therefore, some methods also have been researched to infer or predict user's intention or visual attention by using different factors. Gaze movements correlate with moving in attention and are considered to be a consequence of optimal resource allocation for top-down model tasks such as visual recognition. On the other hand, for the bottom-up model, the features of scene images which human looking at are thought to be the factors for visual attention. Till now, it recognized that the method by computing saliency map is effective for visual attention recognition. Saliency maps proposed by Itti, are often built on the assumption that beforehand features as opposed to objects themselves drive attention. In order to infer or predict user's intention based on gaze movements and saliency maps effectively, it is necessary to study on the approach during the data processing.

On the other hand, in the last years, the applications for developing nursing-care systems for disabled people also have been increased, and a great advance has been produced in the communication systems between humans and machines. One of the great steps is interactive application for the disabled people enabling them self-controlled mobility without external help. But unavoidable problems still exist in these systems for two reasons: their limited

mobility and restricted functionality. Because of their kinematic constraints, conventional wheelchairs are hardly suitable to move within packed rooms. An increase of the mobility can be reached by conducting of an interactive system for omnidirectional wheelchair.

In this research, our final target is to develop an interactive system for omnidirectional wheelchair based on eye tracking. We first propose an eye tracking system to control a humanoid robot, concerning the control method and how to reflect user's operation intention or object. In this system, we address the problem of recognizing the operation intentions of disabled people to a head mounted device with eye tracking function. Neural network is employed in the calibration and tracking process for improving the accuracy and flexibility of the system. The experiment by using a humanoid robot is performed based on the result of eye tracking and executing an object recognition function as the same time. After confirming by user, the robot is controlled with an assistant task for user precisely.

Second, we propose a method to infer user's intention by considering the subjective factors. In this part, the frequency and time of user's gaze appeared in a certain region are considered as the inference basis. They are employed as the input of a fuzzy system to achieve intention recognition. The fuzzy rule used in this part is generated based on the average value of each factor. And an initial set of possible intention regions are also found by object recognition.

We also propose two methods to infer user's intention by considering the non-subjective factors. In extension of previous models of saliency-based visual attention, we propose two methods of bottom-up salient region selection, which estimates the approximate extent of region attended by fuzzy inference and fuzzy neural network (FNN). In these both two methods, the color, intensity and orientation feature maps of an image are employed as the inputs. In traditional method, saliency maps are obtained by combining feature maps. A lot of researches on saliency map are getting some features of image and combining them by simply sum in mathematics. The method of simple sum of them gives them the same importance at the same time. Its weakness is that features' importance in decision process of saliency map cannot be reflected. In order to solve this problem, fuzzy inference rules are making according to the importance of each feature based on expert's experience. But the method is only suitable for specific images, which means is not universal. Therefore,

by using FNN, the importance of all features can be reflected in fuzzy rule with the human decision making model by the conceptual framework of fuzzy logic. In the system with FNN, a McGill calibrated color Image Database is used as the sample data for training.

Finally, to evaluate the effectiveness of proposed methods, inference results of user's intention by using both eye tracking and saliency maps are used to control an omnidirectional wheelchair. Furthermore, we use images gotten from a camera as the input image of the system in order to provide an omnidirectional view to the user. The whole final system is composed by three parts, which are eye tracking, intention recognition and control of omnidirectional wheelchair.

Contents

Chapter 1

Introduction

In this chapter, the background of this research is described. We also analyze the concerned problems still exist. And following, an interactive system for omnidirectional wheelchair by using intention recognition based on eye tracking and saliency maps is introduced as our research target. At last, the structure of the thesis is described.

1.1 Background

According an investigation of Ministry of Health, Labour and Welfare of Japan in 2011, there are about 3.94 million physically disabled people. This means that more and more people of its population are facing experience of functional problems even in their daily lives. Therefore, a lot of researchers and research groups are focus on the related researches to solve the problems such as communication and mobility to improve their quality of lives, ability to live independently and to integrate into society. Relevant government departments have also given some great supports to this work.

In recent years, there have been an increase in the development of assistive technology for people with several disabilities, and great strides have been made in communication systems between humans and machines[1]. Through this assistive technology, disabled people can express their intentions, get more information from the outside and even get their purposes effectively implemented by executing agency such as machines and robots. Most related researches focused on the controlling of wheelchair by disabled people without external help. But a traditional wheelchair is difficult to operate in narrow or crowded areas

such as bathrooms, offices and hospitals. Moreover, it is difficult for a beginner to navigate a wheelchair through a narrow or a complex space because it is always necessary to keep in mind the width of the wheelchair and the surrounding environment[2]. Therefore, the omnidirectional wheelchair with intelligent navigation guidance is researched widely to solve this issue.

However, the interaction between humans and operation object is achieved by traditional methods, for instance, keyboard, mouse, joystick or tactile screen and so on[3, 4, 5]. Most of these systems are manual. At the same time, automatic ones also exist such as voice, gesture and so on[6, 7].

Eye tracking is an important component in many applications including human computer interaction, virtual reality, driver assistance, and diagnosis or early screening of some health problems[8, 9, 10, 11]. Recently, eye tracking technology, with its special advantages of applying to such as ALS patients, is useful as a solution for disabilities or patients to interact with computer and others. Different from other methods mentioned above, the interaction by eye tracking can make user feel more convenient and direct, especially for those who need interact with computer but cannot use keyboard and mouse directly. For instance, the ALS patients, who ultimately lose the ability to initiate and control all voluntary movement, only the muscles responsible for eye movement and eye tracking based interaction is an only optimum way.

But at the same time, although the gaze position can be estimated well in different ways, there are still some problems exist[12, 13]. The most important one is that the saccades phenomenon when we looking at or for something. When the eyes stop to focus it is called a fixation and the movements between these fixations are called saccades. If gaze is used to interact with computer or control an agent directly, user's intention is difficult to recognition because the saccades. Especially for an executive unit, the high speed gaze movement leading to a mismatch between control command and mechanical operation speed. Therefore, it is worth to study on how to recognize user's intention effectively.

On the other hand, recently, the intention recognition, recognizing the intention of a user or an agent by analyzing their actions or changes of state, is becoming an important issue in various research fields of intelligent systems[14, 15, 16, 17]. And actually, it is obvious

that attention takes place in the present, and intention concerns itself with the future[18]. Selective visual attention provides the brain with a mechanism of focusing computational resources on one object at a time, either driven by low-level image properties (bottom-up attention) or based on a specific task (top-down attention). Moving the focus of attention to locations one by one enables sequential recognition of objects at these locations[19].

Previous theory holds that spatial attention results from weaker activation of the same brain circuitry that drives saccadic eye movements[20]. This suggests that if an eye movement is made to a particular location, attention will arrive first and cannot be sent elsewhere. Indeed, the costs and benefits of intentional cueing, or indicating a search target's future location, can be eliminated by requiring saccades to other locations[21]. Actually, we can consider that there are two parts of factors guide attention: subjective factors and non-subjective factors. For example, when subjects with special interests at something, finding something or have a special task, the attention will depend on the subjective factors. And at the same time, attention also driven by the non-subjective factors such as low-level features such as contrast, color, orientation, flicker, or motion of the object or scene[22]. Most attention models are based on a saliency map[23], which is computed from local feature contrasts, for salient locations in the order of decreasing saliency. In other words, saliency of a image is the point or region where the specific feature value is bigger comparing with the rest. Here the specific feature is defined by researchers according experiment situation. And it can be color, intensity, orientation, and so on. Saliency of individual feature has an indirect effect on human attention.

1.2 Research Motivations

Eye tracking technology can be generally divided into two types: remote camera based method[4] and wearable device based method[24]. Except the ability of measuring eye movements, another important factor for the system is that it should be easy to be used and do not make the user feel uncomfortable at the same time. The remote camera based method achieves by setting one or more cameras on or under the computer screen to capture user's eye image for further image processing and calculation. The other method also uses

camera to obtain user's eye image, while the only difference is realized by attaching the camera to a glass or a helmet. And both of the two methods need a calibration process before ready to use. In other words, the system is valid only for the subject who done the calibration process. And at the same time, because the calibration process, in essence, is a coordinates mapping process between user's view plane and pupil moveable plane by geometric transformation, errors will occur during the measuring of eye movements when user move his/her head. So user will asked to try to remain standstill during experiment, which will lead to nervous and uncomfortable. Hence, in order to solve this problem, finding a proper method which can improve fault tolerance of the system is a crucial issue.

On the other hand, as mentioned in Section 1.1, it is difficult to interact with computer or control an agent by gaze directly because of the high speed feature of eye movements. To solve this problem, some related researches focus on the related application development. For example, use dialog boxes, toolbar buttons, menu items and so on as interfaces. But several questions arise in actual operation. Indeed, changing blindness experiments[25] suggest that after looking at pictures of complex natural scenes, we retain information about only the overall gist of the scene and a handful of objects. The experiments show that we generally miss differences between two versions of the same picture, where differences have been introduced by photo editing, if changes are restricted to objects inessential to the overall meaning. This is related to a basic feature of human vision. We move our eyes about three times a second in a pattern of pause and rapid movement: fixation and saccade. In other words, it is also difficult for us to fix at a same position in order to interact with computer by the interface mentioned above. Therefore, if we can obtain user's purpose or intention by analyzing not only the real time gaze positions by the period features of them with other factors, the problem will solve effectively.

Actually, user's intention is guided not only by the subjective factors but also the non-subjective ones such as the role of low-level features, contrast, color, orientation for example. To take non-subjective factors into the recognition or prediction of intention, most models are based on a saliency map and a dynamical process for visiting saliency maxima[23]. Filtering the input image with kernels saliency information of early visual mechanisms generates feature maps at various spatial scales. These are then combined

into a single saliency map, which encodes the probability that an image location will be attended. According to the research of [26], the saliency map is obtained by summing of few feature maps of image most of the time. However the method has a weakness ignoring of the features importance in the decision process of saliency map. Specifically in the case of the feature whose value is low but plays an important role in the process. Hence, it is becoming an important issue to find a method reflecting the importance of each feature properly.

Basing on these backgrounds, in order to combine the eye tracking, intention recognition and control of omnidirectional wheelchair so as to construct an intelligent interactive system, in this thesis we aim at three points as follows.

- We propose methods to improve the accuracy of eye tracking by changing the calibration method and using soft computing method. Both two modes of eye tracking system, head-mounted mode and remote mode, are used in our research and by analyzing the eye tracking results, the relationship between eye movements and intention is explored.

- The prediction methods of the visual attention region inferred by using fuzzy inference and fuzzy neural network (FNN) after extracting and computing of images feature maps and saliency maps are proposed. Finally, eye tracking results are also factored in this process.

- Based on the results of intention recognition, we propose an interactive system for omnidirectional wheelchair by using an omnidirectional wheelchair. The system can provide an omnidirectional scene of surrounding environment and have omnidirectional mobility.

1.3 Structure of This Thesis

This thesis is composed by 7 chapters and organized as following.

In chapter 1, the research background and objectives of this thesis are introduced.

In chapter 2, the necessary background knowledge of eye tracking needed to comprehend the research is provided. We summarize different techniques for eye tracking. Based on their geometric and photometric properties, the techniques can be classified in three categories: contact lens based, electrooculogram based and video-based. Alternative techniques may exploit motion and symmetry. We also describe that the active IR illumination may be employed by various techniques.

In chapter 3, we describe the review with references to those areas which touch upon intention, either researching it directly or using it in service of other branches of study. After establishing the importance and applications of the study of intention, we turn to inquire into the different meanings of the term, since it is used in different contexts, and not always with the same meaning. Various characteristics of intention can be the potential candidates for solving the problem of intention recognition. Several intention recognition models are described. A typical architecture for human-machine interaction with intention is also illustrated. By analyzing the research status of intention recognition, we find that there are two categories for it: eye movements based and visual attention based. Both of the two categories and the typical application of them are also introduced.

In chapter 4, two models of eye tracking system used in this research are described respectively. The first one is wearable device based eye tracking system. And in this research, the device is built by us. For this device, the structures of hardware and software are explained. And in the software part, we focus on the calibration process of system and use neural network instead of the traditional calibration method. Experiment also designed and proceeded to verify the effect of proposed method, which shows that a better performance can be obtained. The second eye tracking system is remote camera based. The hardware specifics are described in detail. And by using this device, a series of experiments have been carried out, which are analysis of gaze order, stagnation map and focus area, to illustrate the relationship between gaze movement and attention.

Chapter 5 is mainly consisted by three parts. In the first part, we proposed an approach for intention recognition based object recognition by using eye tracking. Firstly, object learning process is carried on by using a humanoid robot. Following, the region set that

user may pay his/her attention on is constructed by object recognition, where the scene image captured by a camera on robot is used as input for image processing. At the same time, user's eye movement is tracked by an eye tracking device. Finally, by comparing the result of object recognition and user's gaze distribution, intention region or object can be decided. Through this experiment we find that the process of object learning brings restrictions of widespread application to the system.

In the second part, by analyzing the models used for visual attention, we find that the top-down model for visual attention is difficult to understand compared with bottom-up model. An effective and important method for visual attention in bottom-up model is by using saliency map. Saliency map in this research is calculated by using color, intensity and orientation feature maps of image. Focusing on this issue, we propose two methods for obtaining of saliency map, comparing with the traditional methods. One of the methods is by using fuzzy inference. Considering the traditional method can not reflect the importance of each feature of input image, however in this method, the importance of feature is embody in fuzzy rules, which obtained by expert experience. The experimental results show that the method can solve the problem. But at the same time, its application may be under restrictions because the fuzzy rules can not be adjusted dynamically.

To improve the method mentioned above, a new method by applying fuzzy neural network is proposed in the third part. The calculation method is same but the importance of all features can be reflected in fuzzy rule with the human decision making model by the conceptual framework of fuzzy logic. The sample data for training is obtained by using an image database. The proposed methods are evaluated experimentally by using eye tracking system and questionnaires completed by participate. At last, the overall procedure for attention prediction by combining both eye tracking and visual attention is also proposed.

In chapter 6, user's intention is recognized by combining the result of eye tracing and attention prediction. An omnidirectional wheelchair is introduced and the specific, especially the kinematics and control method are described at the same time. The directions of movement and turning of the omnidirectional wheelchair can be achieved by setting appropriate angular and speed of its wheels. By applying the intention recognition result into the controlling of an omnidirectional wheelchair, an interactive system for omnidirectional

wheelchair based on eye tracking is constructed. Receiver operating characteristic method is employed in the combination process of gaze distribution map and saliency map in order to obtain credible results. Experiments also have been carried out to evaluate the system. The conclusions and future work are presented in chapter 7.

Chapter 2

Eye Tracking Techniques

2.1 Introduction

HCI (human-computer interaction) has been a major area of research in computer science, human factors, engineering psychology and closely related disciplines. And current interaction methods tend to be one side, with the bandwidth from the computer to the user far greater than that from user to computer. A fast and effortless mode of communication from a user to a computer would help redress this imbalance.

Eye tracking is a technique whereby an individuals eye movements are measured so that the researcher knows both where a person is looking at any given time and the sequence in which their eyes are shifting from one location to another. Tracking peoples eye movements can help HCI researchers understand visual and display-based information processing and the factors that may impact upon the usability of system interfaces. In this way, eye-movement recordings can provide an objective source of interface-evaluation data that can inform the design of improved interfaces. Eye movements can also be captured and used as control signals to enable people to interact with interfaces directly without the need for mouse or keyboard input, which can be a major advantage for certain populations of users such as disabled individuals. We begin this chapter with an overview of eye-tracking technology, and progress toward a detailed discussion of the use of eye tracking in HCI and usability research.

Research in eye detection and tracking focuses on two areas: eye localization in the image and gaze estimation. There are three aspects of eye detection. One is to detect the

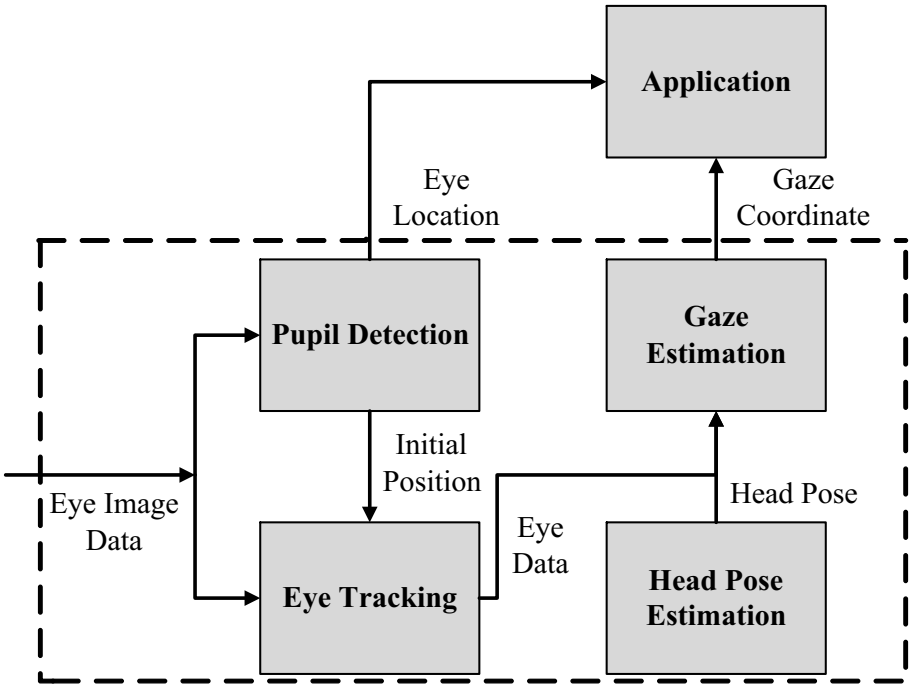


Figure 2.1: Components of Video-based Eye Tracking System

existence of eyes, another is to accurately interpret eye positions in the images, and finally, for video images, the detected eyes are tracked from frame to frame. The eye position is commonly measured using the pupil or iris center. A general overview of the components of eye tracking system is shown in Fig.2.1. Video systems obtain information from one or more cameras (Image data). The eye location in the image is detected and is either used directly in the application or subsequently tracked over frames. Based on the information obtained from the eye region and possibly head pose, the direction of gaze can be estimated. This information is then used by gaze-based applications, for example, moving the cursor on the screen.

2.2 Electrical Based Eye Tracking

Generally, the eye tracking devices measure/determine the eye ball position in several ways by analog techniques that can be classified in two categories: contact lens based and

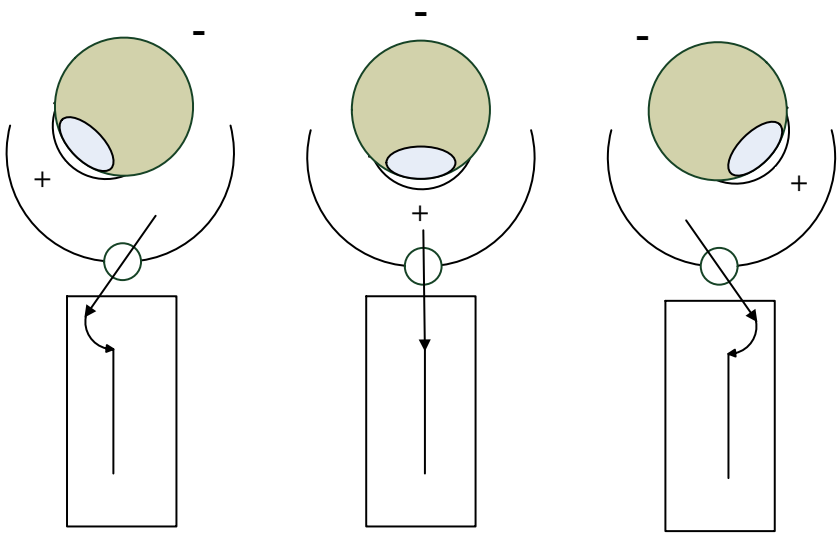


Figure 2.2: Eye Ball Polarization in EOG[1]

electrooculogram based.

The first category includes invasive eye tracking that use contact lens with mirrors[27] or magnetic search coil[28]. The eye tracking that uses contact lens with mirrors implies an entire process of attaching the lens to eye ball and the experiment can last only a short period of time (measured in minutes). The eye tracking with magnetic search coil requires two soft contact lens and between a coil of wire with 13 mm diameter. These eye tracking were used specially used by the scientists for research of physiology and dynamic of eye movements. Despite the vast improvements and the accuracy obtained, the systems were not widespread because of invasive process of attaching the lens and because the head had to be kept still in order not to affect the measurements.

The eye tracking from second category measure the eye balls biopotentials using electrodes placed near the eye. It is also called electrooculography (EOG) method. This potential can be considered as a steady electrical dipole with a negative pole at the fundus and a positive pole at the cornea, as shown in Fig.2.2. The standing potential in the eye can thus be estimated by measuring the voltage induced across a system of electrodes placed around the eyes as the eye gaze changes, thus obtaining the EOG (measurement of the electric signal of the ocular dipole). The electrooculogram is captured by five electrodes

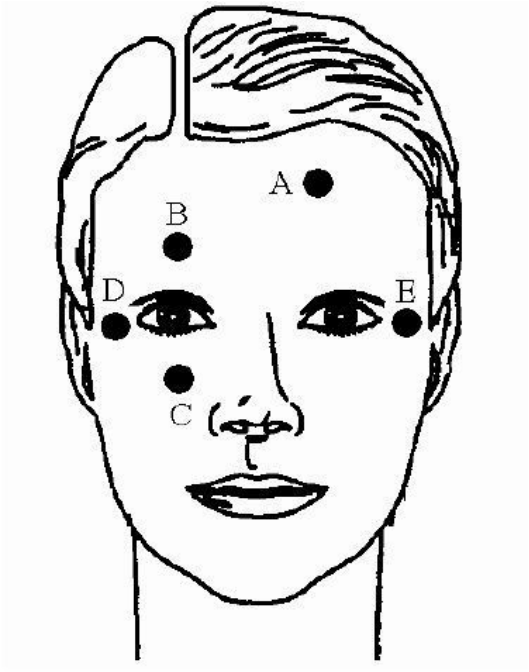
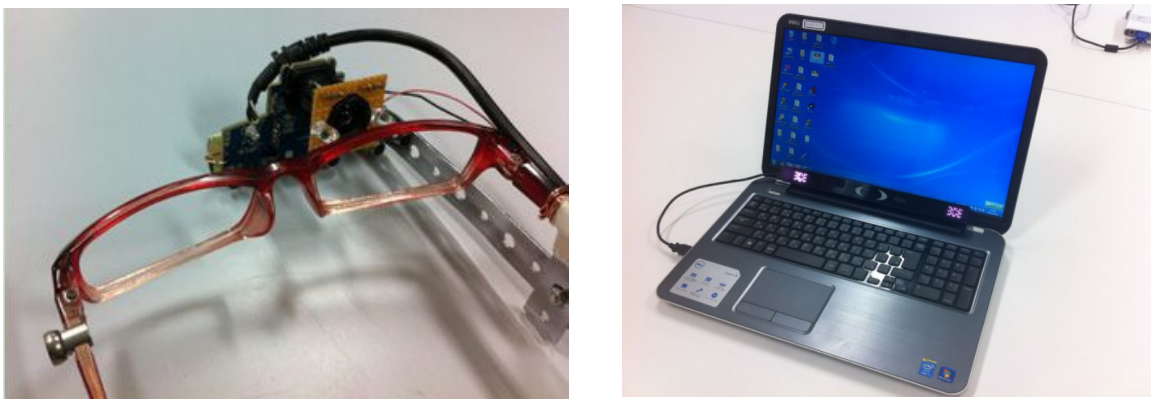


Figure 2.3: Electrode Placement in EOG[1]

placed around the eyes, as shown in Fig.2.3. The EOG signals are obtained by placing two electrodes to the right and left of the outer canthi (D-E) to detect horizontal movement and another pair above and below the eye (B-C) to detect vertical movement. A reference electrode is placed on the forehead (A). The EOG signal changes approximately 20 micro volts for each degree of eye movement. This is possible to determine the eye positions and used in human computer interaction. The disadvantages are the costs of signals amplifiers and the presence of electrodes on subject face.

2.3 Video-Based Eye Tracking

Beside by analog techniques, an another method for eye tracking uses a video camera to track the position of the eye. Different from electrical based eye tracking model, this method measures eye movements by using camera in front of human. The camera gets human's eye image first, and by image processing of the eye image, pupil center and gaze position can be calculated. The model has the advantage that no attachments needed to put on human, which will make participant feeling little comfortable. Another advantage is



(a) Head-mounted Eye Tracking System (b) Remote Eye Tracking System

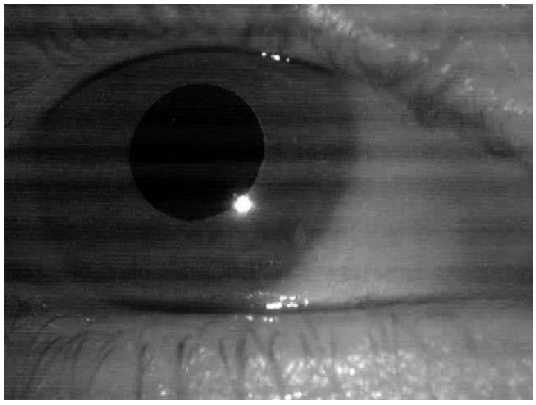
Figure 2.4: Two Types of Video-Based Eye Tracking

that the higher resolution the camera has, more precise measuring result can be obtained.

2.3.1 Two Types of Video-Based Eye Tracking System

Video-based eye tracking systems can be divided into remote and head-mounted systems (see Fig.2.4). Each type of system has its respective advantages. Both visible spectrum and infrared-spectrum imaging techniques have been applied in the context of remote video-based eye tracking. The single most attractive reason for using a remote eye-tracking system is that its use can be completely unobtrusive. However, a limitation of a remote system is that it can only track eye movements when the user is within a relatively concerned area of operation. And the accuracy of remote eye-tracking systems is usually worse than the head-mounted eye-tracking systems. Stereo cameras can be applied to achieve better eye-tracking accuracy[29, 30]. The design of remote eye-tracking systems must consider the three way trade-off between cost, flexibility and quality. For example, the flexibility to track eye movements over a wide area can be improved by using a pan-tilt camera, but such cameras are quite expensive. Furthermore, the quality of eye tracking can be improved by capturing a high-resolution image of the eye using a zoom camera[31], with the trade-off of a reduced operational area and higher cost. Although, there is a number of promising remote eye tracking approaches[32], it currently appears that a head-mounted system has a greater potential to achieve a reasonable compromise between all of these factors.

(a) Eye Image Captured with Visible Spectrum Imaging



(b) Eye Image Captured with Active IR Illumination

Figure 2.5: Eyeball Images

2.3.2 Visible Image and Active IR Illumination Image

Two types of imaging approaches are commonly used in eye tracking, visible image and active IR illumination image[33], as shown in Fig.2.5. The three most relevant features of the eye are the pupil-the aperture that lets light into the eyeball, the iris-the colored muscle group that controls the diameter of the pupil, and the sclera, the white protective tissue that covers the remainder of the eye.

Visible spectrum imaging is a passive approach that captures ambient light reflected from the eyeball. In these images, it is often the case that the best feature to track is the contour between the iris and the sclera known as the limbus. Visible spectrum eye tracking is complicated by the fact that uncontrolled ambient light is used as the source, which can contain multiple specular and diffuse components. Infrared imaging eliminates uncontrolled specular reflection by actively illuminating the eye with a uniform and controlled infrared light not perceivable by the user. A further benefit of infrared imaging is that the pupil, rather than the limbus, is the strongest feature contour in the image. Both the sclera and the iris strongly reflect infrared light while only the sclera strongly reflects visible light. Tracking the pupil contour is preferable given that the pupil contour is smaller and more sharply defined than the limbus. Furthermore, due to its size, the pupil is less likely to be occluded by the eye lids. The primary disadvantage of infrared imaging techniques is that

they cannot be used outdoors during daytime due to the ambient infrared illumination.

Infrared eye tracking typically utilizes either a bright-pupil, dark-pupil technique or both. The bright-pupil technique illuminates the eyeball with a source that is on or very near the axis of the camera[34]. The result of such illumination is that the pupil is clearly demarcated as a bright region due to the photo reflective nature of the back of the eye. Dark-pupil techniques illuminate the eye with an off-axis source such that the pupil is the darkest region in the image. While the sclera, iris and eye lids all reflect relatively more illumination. In either method, the first-surface specular reflection of the illumination source of the cornea (the outer-most optical element of the eye) is also visible. The vector between the pupil center and the corneal reflection center is typically used as the dependent measure rather than the pupil center alone. This is because the vector difference is less sensitive to slippage of the head gear-both the camera and the source move simultaneously.

Chapter 3

Traditional Researches on Intention Recognition

3.1 Introduction

Human intention recognition is crucial for an efficient human computer interaction. Recently, intention modeling and recognition are being perceived in psychology and cognitive science to create a new paradigm of human computer interface (HCI) and human robot interaction (HRI)[35, 36]. Human intention can be explicit or implicit in nature. Generally, humans express their intention explicitly through facial expressions, speech, and hand gesture. In HCI and HRI, the user intention can be explicitly conveyed through a keyboard and a computer mouse[37], which can be easily interpreted. However, the human's implicit intention is vague and is difficult to understand. Interpreting the user's implicit intention, which contains valuable information in addition to the explicit intention, is vital in developing an efficient human computer interactive system.

Several philosophers like Kellogg[38] and Dennett[39] discussed the role of intention in theories of consciousness, planned action, rationality, and intelligence. As mentioned, Kellogg considers intention as the main attitude that directs future planning. Dennett on the other hand argues that the ascription of intent is a perfectly reasonable way of predicting and describing the behavior of systems that are complex enough to avoid explanation from other stances[40]. He offers three stances that can be used to explain and predict the behavior of the system: the physical stance that operates directly upon knowledge of physical composition and of the laws of physics, the design stance that deals with the

purpose the system is designed for, and the intentional stance where a system is assumed to have certain beliefs and desires and it behaves in such a way to further its goals in light of its beliefs.

Human interaction usually requires continuous and complex intention recognition. In simple conversations, for example, humans try to predict the future direction of the conversation and the reaction of the other person by recognizing the intent manifested in the conversation. Further, in a promenade, two persons mutually coordinate their steps and moves by predicting the intention of each other. Often intention recognition by humans is performed subliminally; conversations unfold and subtle queues are reacted to without much conscious thought given to the underlying motives of the other[40].

Intention recognition can be seen as a substitute or complimentary to reliable and extensive communication which is a prerequisite for coordination and cooperation. If agents are able to express their intent clearly and honestly then intention recognition reduces to communication. But since not all agents are explicitly aware of their intentions or since communication can be a burden on agents (different designs, different levels of intelligence, or heterogeneous ontologies) then intention recognition becomes essential. Furthermore, it is sought to have a natural interaction in human-machine cooperation; natural to a level that resembles human-human interaction.

Intention recognition is defined, in general terms, as the process of becoming aware of the intention of another agent. More technically, it can be defined as the problem of inferring an agents intention through its actions and their effects in the environment. It lies accordingly in the boundary between perception and cognition.

The ultimate goal of related researches is to incorporate the intention recognition module into human-machine interaction to achieve compliant and intelligent interaction. For this, an architecture for human-machine interaction (HMI) is shown in Fig.3.1[63]. Classical modules as human-machine interface and supervision and control can be integrated in this architecture. Although one of the goals of intention recognition is to minimize the need for traditional interfaces, it is still possible to augment certain interaction aspects with visual and auditory interfaces.

In the architecture in Fig.3.1, the processes in the world are divided to human processes,

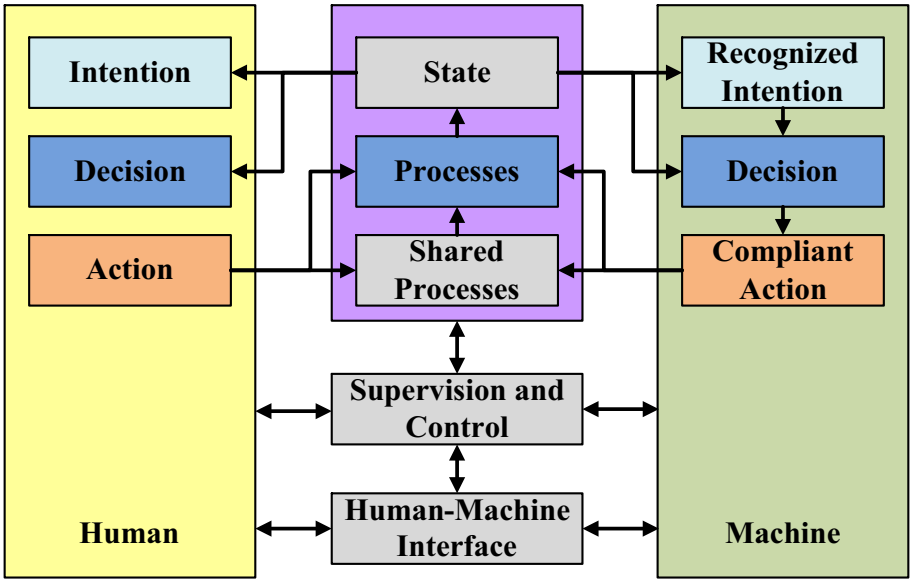


Figure 3.1: Architecture for Human-Machine Interaction

machine processes, and shared processes. The effect of operating these processes can be partially or completely sensed by the human and the machine as world states. The human uses the sensed information to further her goals by selecting actions to achieve her desired states. On the other hand, the machine which has a human-task model, employs the sensed states to recognize the human intention and accordingly to select its action in a compliant fashion with the human intentions.

3.2 Intention Recognition Based on Eye Tracking

In humans, the eye movements are the essential motor movements that are controlled by the human cognitive system[42, 43, 44]. In other words, the eye movements and its position are not random but directly related with the visual information present in the scene and provide a rich window into the human's sensory processing, intentions and thoughts. Therefore, being a window to the mind, the eye and its movements are tightly coupled with human cognitive processes. Therefore, in humans, when viewing a visual scene, different implicit intentions result in different eye movement patterns. Hence, the eyeball movement patterns can be considered as possible factors for recognizing the human's implicit intention. The notion to take advantage of the information present in the eye-gaze leads to

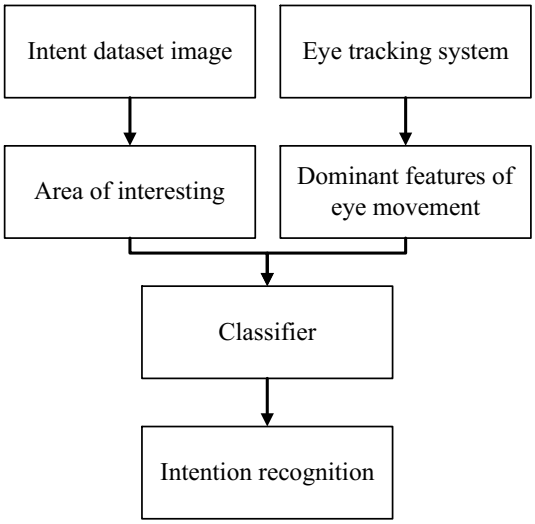


Figure 3.2: A Model for Intention Recognition Based on Eye Tracking

the development of efficient eye tracking equipment which attracted many researchers in human-computer interaction. In [45], the authors present an eye-typing interface based on eye fixation tracing, an automated method to map the eye movements into a process model prediction using hidden Markov model. In [46], the authors combine eye tracking and object recognition to recognize user's intention. In [47], the authors develop an eye based activity recognition (EAR) system using a support vector machine (SVM) and eye movements such as fixations, saccades and eye blinks. In addition to the eyeball movements, the pupil size variation has been studied in relation to cognitive processing and visual information. The pupil size can be used as a measure of the human attention[48, 49].

A model for intention recognition based on eye tracking is as shown in Fig.3.2.

In Fig.3.2, an eye tracking system is used to measure human eyeball movement patterns and the pupil size variation for a given visual stimuli. In the images, which are presented as visual stimuli, areas of interest (AOI) is preset to analyze the eyeball movement pattern by visual attention models and the pupil size variation based on the intention of the subject under consideration. Dominant features of eye movement such as fixation length, fixation count and pupil size variation can be used to classify the subject's implicit intention into navigational and informational intent. Based on the fixation length and fixation count in each AOI of a given input stimulus image and the pupil size variation, different classifiers can be constructed to differentiate the subject's intent into navigational and informational

intent.

Right now, by using eye tracking system, dominant features of eye movement can be measuring effectively. But one problem for this model is that how to decide the AOI. Traditional methods achieve it by using object recognition, which learning process is needed beforehand. This also limits the widespread application of intention recognition.

3.3 Visual Attention Models

Visual attention models aim to predict the attentional behavior of human observers when viewing a visual scene. Generally, these models are not able to predict the sequential order of human fixations, the scan path, but are limited to predicting the locations and objects that humans focus on[50].

Many visual attention models were inspired by early works such as the feature integration theory by Treisman and Gelade[51], the guided search by Wolfe et al.[52], the neural-based architecture by Koch and Ullman[53]. Especially the latter model constituted a theoretical basis for biologically plausible models incorporating characteristics of the human visual system (HVS) known to contribute to visual attention, such as multiple-scale processing, contrast sensitivity, and center surround processing. Probably the most widely used bottom-up visual attention model following this paradigm is the one by Itti[23], which is based on the neuronal architecture of the early visual system, where multiple-scale image features are combined into a topographical saliency map. Other visual attention models based on Bayesian[54], information theoretic[55], or statistical approaches[56].

Only few models thus far have focused on top-down attentional processes[57], mainly because they are relatively less well understood compared to bottom-up attentional processes.

As saliency map is an important method for visual attention, a typical model for saliency map is proposed by Itti[26], as shown in Fig.3.3. Briefly, the model in Fig.3.3 includes two color channels (blue/yellow and red/green), one intensity channel, and four orientation channels (0°, 45°, 90°, 135°). Raw maps of nine spatial scales (0-8) are created using dyadic Gaussian pyramids. Six center-surround difference maps are then constructed as

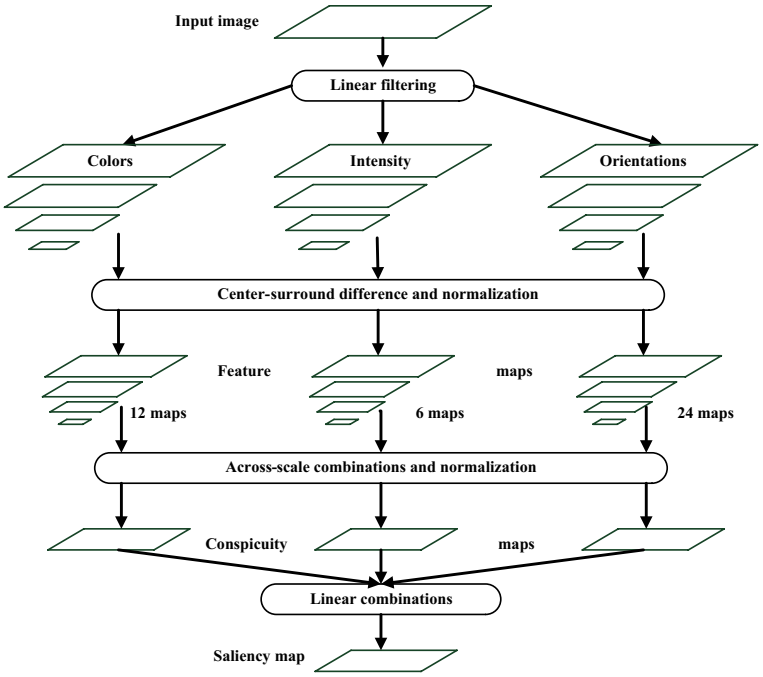


Figure 3.3: A Typical Model for Saliency Map Proposed by Itti

point-wise differences across pyramid scales to capture local contrasts (center level c = {2, 3, 4}, surround level s = c + δ, where δ = {2, 3}). A single conspicuity map for each of the color, intensity, and orientation feature channels is built through across-scale addition of the center-surround difference maps and is represented at scale 4.

Chapter 4

Proposed Eye Tracking System

In order to estimate gaze based on image analysis, it is common that gaze location is estimated with pupil location. The previously published pupil detection methods can be divided into two modes: the wearable device based mode[58, 59] and the remote camera based mode[60, 61]. In this research, we use both two modes in our experiments. In this section, the wearable device based eye tracking system is made by ourselves while the remote camera based one is produced by DITECT[62].

4.1 Wearable Device Based Eye Tracking

In this research, we design an eye tracking system with the features of low cost, easy to use and high accuracy. To raising the accuracy of the system, we propose a new calibration method based on neural networks.

4.1.1 Proposed Eye Tracking Device

The device is a head unit. It is equipped with a camera in order to capture the image of the right eye. The head unit is a glass frame in this research and sends all images to the computer.

As shown in Fig.4.1, the eye tracking hardware made by us is a wearable device including an eye capture camera attached with NIR LED. Detailed specifications of the proposed prototype device are shown in Table 4.1. In order to make tracking the eye much easier, we illuminate the eye with IR light and observe it through an IR sensitive camera with a

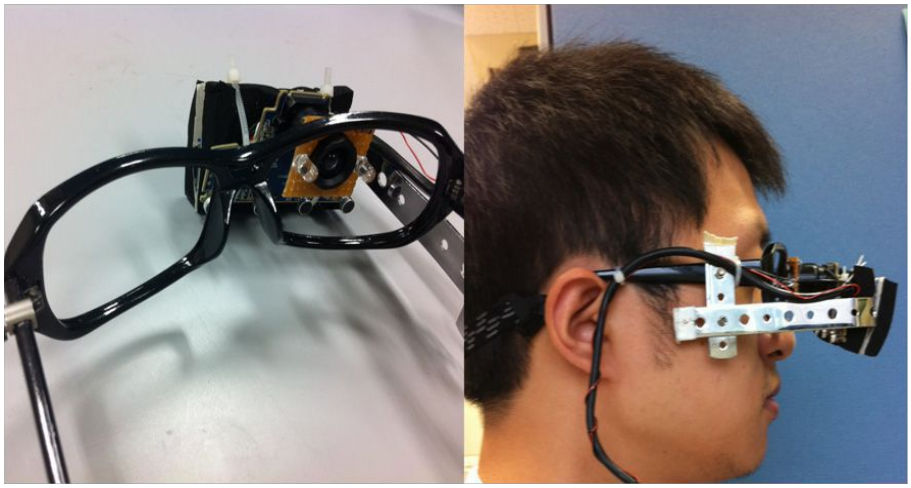


Figure 4.1: Proposed Eye Tracking Prototype Device

Table 4.1: Specifications of Prototype Device

CCD camera	Spatial resolution	640×480 pixels
	Frame rate	60 FPS
	Lens focus	fixed
NIR LED	Wave length	940nm
	Luminous intensity	40mW/Sr
	Angle	5deg

visible light filter. After doing this the iris of the eye turns completely white and the pupil stands out as a high-contrast black dot. We also have investigated the price of eye tracking system sold in the market and found that this price is low.

4.1.2 Pupil Center Detection

The pupil center detection is the first part of an eye tracking system, the most important part at the same time[63]. In this paper, we detect user's pupil center through eye image processing. The schematic diagram of process flow is shown in Fig.4.2.

At first step we capture the eye image by CCD camera and process the image binary, as shown in Fig.4.3. In step 2, the contrast of the image has been increased in order to make the detection process easier. Although the mathematical transformation is a larger change, it is generally not apparent in the image. Then the program will try to find any blobs

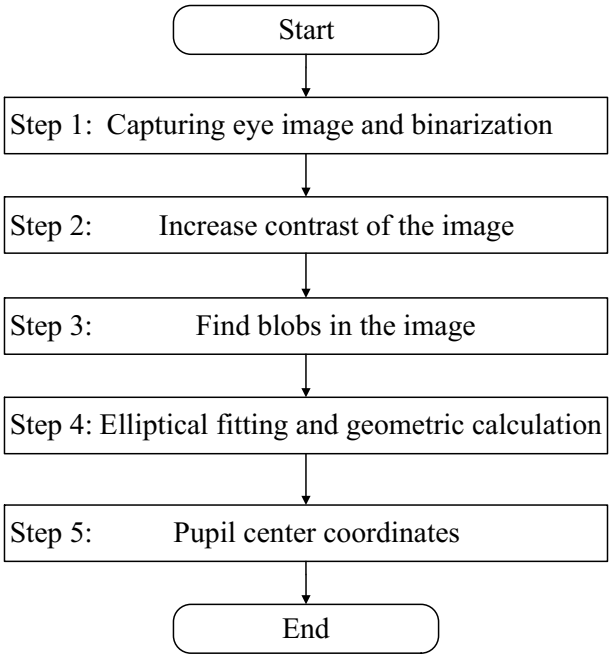


Figure 4.2: Algorithm Flow of Pupil Center Detection

existing in the image and record the feature points of the optimal one after filter in step 3. In step 4, by utilizing Sklansky algorithm[64], the convex shape of these feature points can be calculated. Finally, the pupil center coordinates can be obtained by calculating of geometric center after elliptical fitting in step 5. Finally the coordinates of pupil center can be obtained, which are shown in image 4 of Fig.4.3.

4.1.3 Gaze Estimation

The primary task of eye tracking system is to estimate user's gaze which is also the foundation of interaction between human and computer by this method. In this research, the gaze estimation has been achieved by using neural network (NN) to improve the robustness and adaptability of the system. In the calibration process, based on the coordinates of pupil center got in Section 3.2.1, a two-input and two-output NN with standard back propagation algorithm is used as shown in Fig.4.4.

Where input P_x, P_y and output G_x, G_y are pupil center coordinates in 2D plane of camera image and user's gaze position coordinates in computer screen, respectively. w_{1i} means the weight value between input node (I_1) and the hidden node H_i. In this paper, we

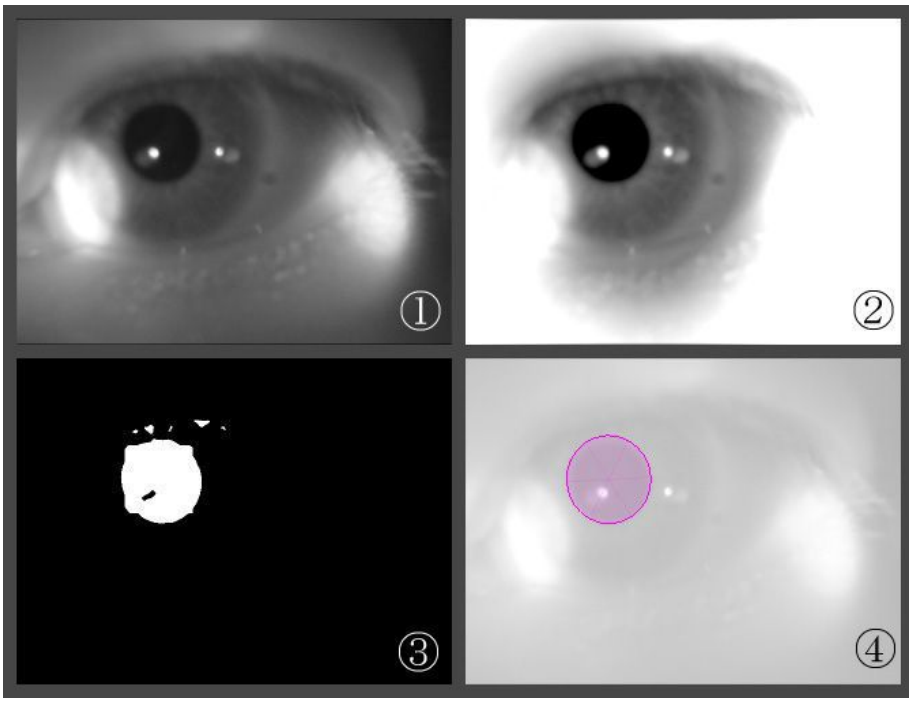


Figure 4.3: Eye Image Capturing and Pupil Center Detection

use the sigmoid function as the transmission function. The parameters of NN are shown in Table 4.2.

In Table 4.2, the desired error is 0.001 at each pixel as units. Because in our program, the inputs of the neural network must in the range of 0 to 1, but actually the data source are screen positions which in our experiment are in the range of 0 to 1366 pixels. The max trails number above is set based on experience. Fig.4.5 shows the mean square errors (MSE) with various choice of neural network's hidden neuron number according to the learning trails of neural network training. From the figure we can find that the training process when the neuron number is nine has a fast convergence rate and a least MSE. Thus, in this paper, we select nine as our neural network's hidden neuron number.

For example, the output value G_x of NN can be calculated as follows:

$$G_x = \frac{1}{(1 + exp(-\sum_{i=1}^{n} \frac{1}{(1+exp(-\sum_{j=1}^{2} I_j w_{ji}))} w_{ij}))} \tag{4.1}$$

Also the output value G_y can be calculated by using the same method. Because the resolution of the computer screen used is 1366×768 pixels, so the range values of G_x and G_y are 0 to 1366 and 0 to 768.

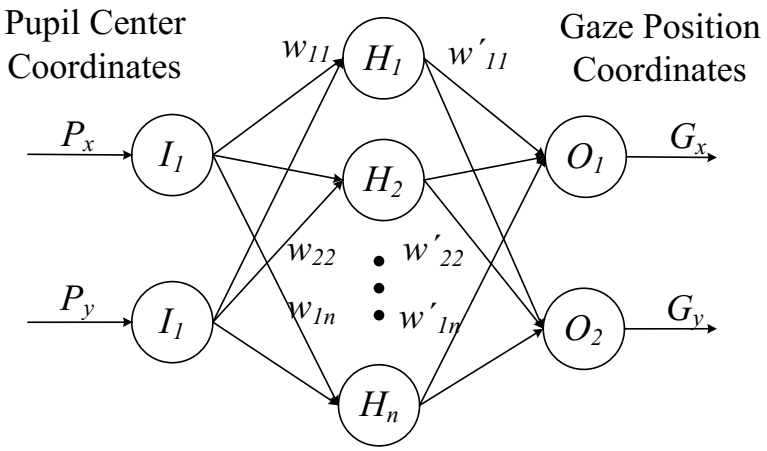


Figure 4.4: Neural Network for Gaze Estimating

Table 4.2: Parameter Setting of Neural Network

Desired error	<0.1%
Maximum trial number	3000
Number of layers	3
Number of hidden neurons	9
Learning rate	0.7
Input neurons	2
Input neurons	2

In the calibration process, developers usually use some designed points such as the calibration points[65]. But sometimes this method may cause a bad calibration result in a repeating experiment. Especially, when a user performs the experiment several times, he/she will move gaze to the next prospective point before present point's calibration finishes. To make the calibration has more universality and reduce the possibility of user's anticipation to the calibration process, a random calibration process is carried out in our experiment. Specifically, the position of each point is given at random so that user's anticipation can be eliminated.

At the same time, the set consisted by all the appeared points must can cover the whole screen plane, as shown in Fig.4.6. Actually, if the mapping function between movable pupil center region and user's view region can be obtained using geometric transformation. However, this method has the problem that mapping function should be changed according

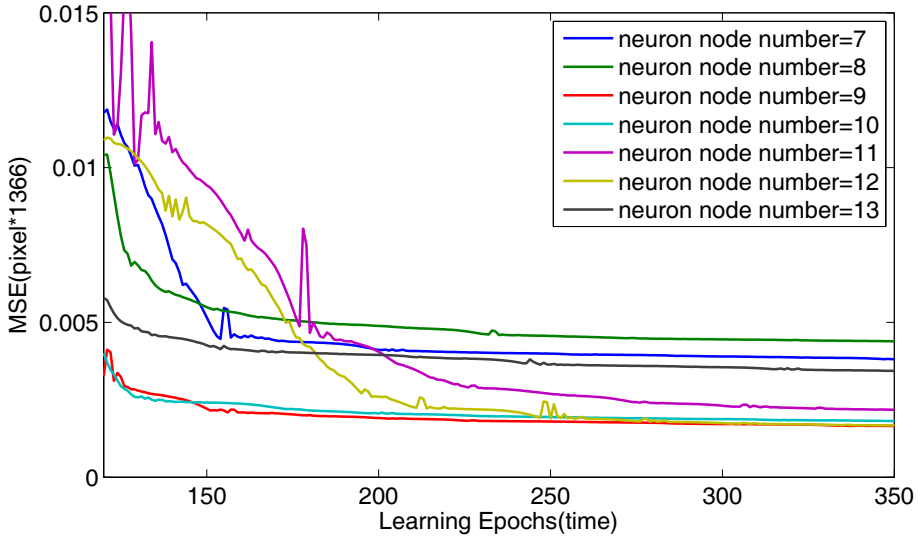


Figure 4.5: Training Processes of NN According to Different Hidden Neuron Nodes Number

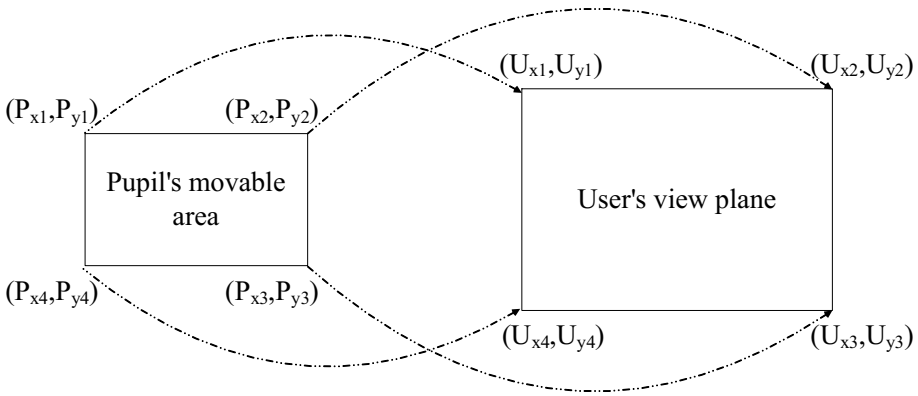


Figure 4.6: Coordinate Mapping Between User's View Plane and Pupil's Movable Area

to the distance (Z) between user and computer screen. This is because the user's view region will change according to Z distance although the movable pupil center region is the same. To solve this problem, neural network is employed in this research. In Fig.4.6, the rectangle on the right side is stand for users view plane in our experiment. And in this paper the view plane is the computer screen.

According to the resolution of the computer screen, in Fig.4.6, (U_{x1}, U_{y1}) is the upper-left corners coordinate (0,0). And so on, (U_{x2}, U_{y2}), (U_{x3}, U_{y3}), (U_{x4}, U_{y4}) stand for the upper-right, lower-right and lower-left corner of the screen, where the coordinate values are (1366,0), (1366,768) and (0,768) respectively. The rectangle on the right side is stand for pupil's movable area. When user looked at the upper-left corner of the screen, where

the coordinate is (U_{x1}, U_{y1}), the coordinate value (P_{x1}, P_{y1}) is (190,144) in the eye image, resolution is 640×480 pixels. And when looked at the other three corners, the coordinates (P_{x2}, P_{y2}), (P_{x3}, P_{y3}) and (P_{x4}, P_{y4}) are (434,148), (440,329), (183,327), respectively. In theory, if the coordinates (P_{x1}, P_{y1}) and (P_{x3}, P_{y3}) are precise, the other two coordinates (P_{x2}, P_{y2}) and (P_{x4}, P_{y4}) should be (440,144) and (190,329). We considered that the users head cannot remain perfectly still in the experiment process cause the slight error. We also think that this is acceptable.

4.1.4 Experimental Results

The experiment is in order to verify the accuracy of the eye tracking system after calibration. And the time of calibration process took 24(1.5×16) seconds. The experiment using the proposed method of eye tracking is conducted on a notebook computer with Intel Core i3-380M CPU, 2 GB RAM and Microsoft Windows 7 operating system. The program is developed in Code Blocks which is an open source IDE and Matlab R2007a. The part of pupil center detection is achieved by using OpenCV and openFrameworks, an open source C++ toolkit.

In the experiment, we used 16 points as the reference points and user is demanded stare at each point for 1.5 seconds. In this process, the usually method is that 16 points appeared in a 4×4 grid which is so-called standard points are prepared and the user is asked to look at the points when they appear, as shown in Fig.4.7, and the coordinates of user's eye gaze estimation position are recorded and used as mapping data together with the appearing position of the standard points.

The order the points appear is according to an "S" type as shown in Fig.4.7. But we found that user's behavior was easy to be a habitual after the experiment carried on for several times and caused a bad calibration result. So we use the method by making the standard points to appear randomly to replace the before one. Similarly, when the standard points appearing randomly, we record the positions coordinates of both standard and estimation ones.

Because the image acquisition speed is 60 FPS, there are about 1440(1.5×60×16) points are used as the input of the neural network and also the same number for the output.

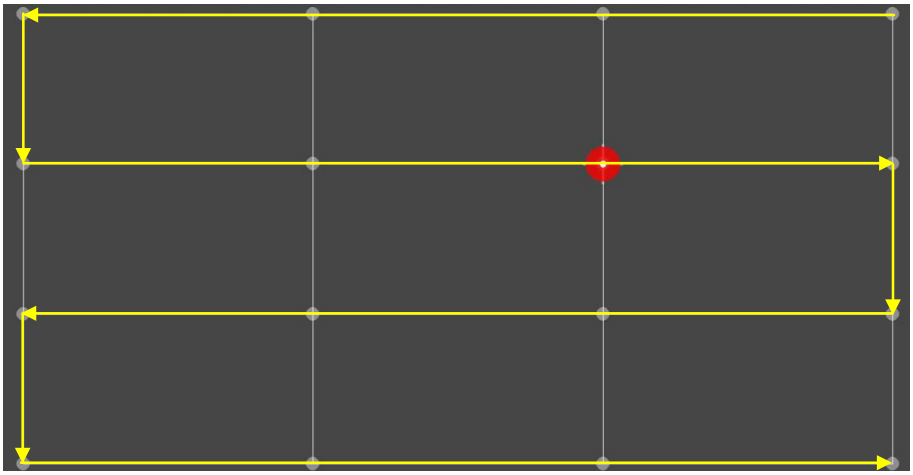


Figure 4.7: Position of Standard Points

Table 4.3: Accuracy of Eye Tracking

	X-axis	Y-axis
Distance(%)	3.24	3.68
Direction(deg)	1.172	0.998

Because the neural network used here is a BP (Back Propagation) method, so the teaching signals are also composed by the above-mentioned coordinates. And after 1000 trails in calibration process, the actual error meets to the desired error 0.001, which is set at Table 4.2.

After calibration, we carry out an experiment to validate the calibration results. In this experiment, 16 points are giving at first as reference points. Next, the user is asked to look at each point in sequence. The position data of reference points and user's gaze at each position then will be recorded at the same time. Then the average axes of gaze positions are calculated and plotted in a figure together with reference points. The distance between each estimated point and reference point are shown in Fig.4.8. The red points in the figure are reference points shown to user after calibration process while the blue ones are users actual gaze positions on computer screen when he/she looked at the red points. In this experiment, the distance between user's eye and computer screen is 45cm. The average error of the results is shown in Table 4.3. The values embody that the error between users actual gaze position coordinate and the intended coordinate is within acceptable range.

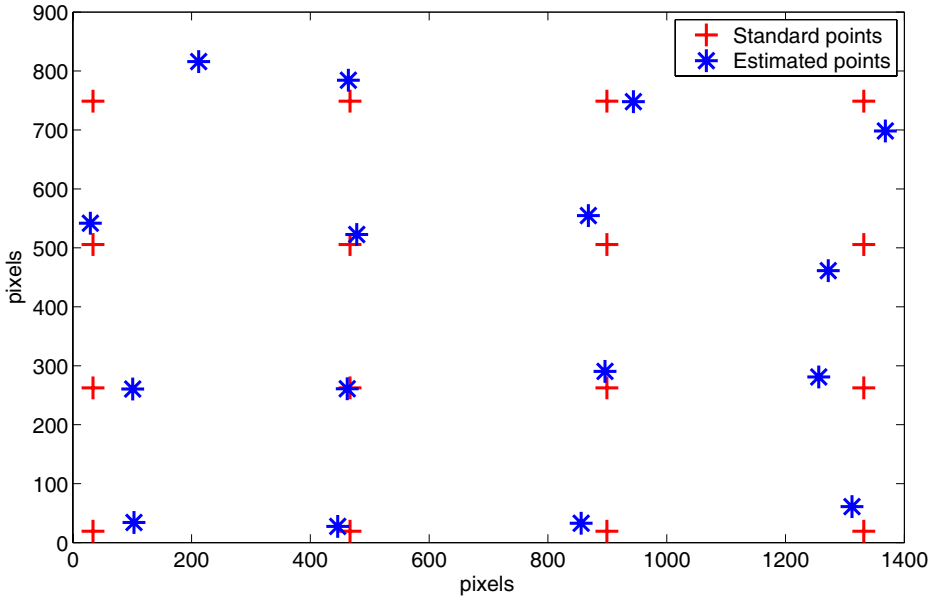


Figure 4.8: Experimental Results after Calibration

In our experiment, the distance between user's eye and computer monitor is 45cm and the resolution of the screen is 1366×768 pixels. The schematic diagram is shown in Fig.4.9. At the same time the width and height of the screen are 28.448cm and 21.336cm, respectively. Thus, one pixel in the computer screen stand for about 0.02cm. For example the average error in pixel is t, the degree error can be calculated as follows.

$$\theta \approx \arctan(0.02t/45) \tag{4.2}$$

4.1.5 Analysis of Gaze Distribution

As mentioned in the section 1.1, the user's visual attention region cannot decided only by the features of an image, because of different users who have different interests. And user's different interests can be reflected well by his/her gaze distribution when looking at something. But the estimation of attention region also cannot be done with only user's gaze position data, because the estimation will become a direct judge system and lose the prediction sense.

According the reasons mentioned above, we make a mathematical statistics of gaze distribution in a period after user looked at the image at first. And as compared with the

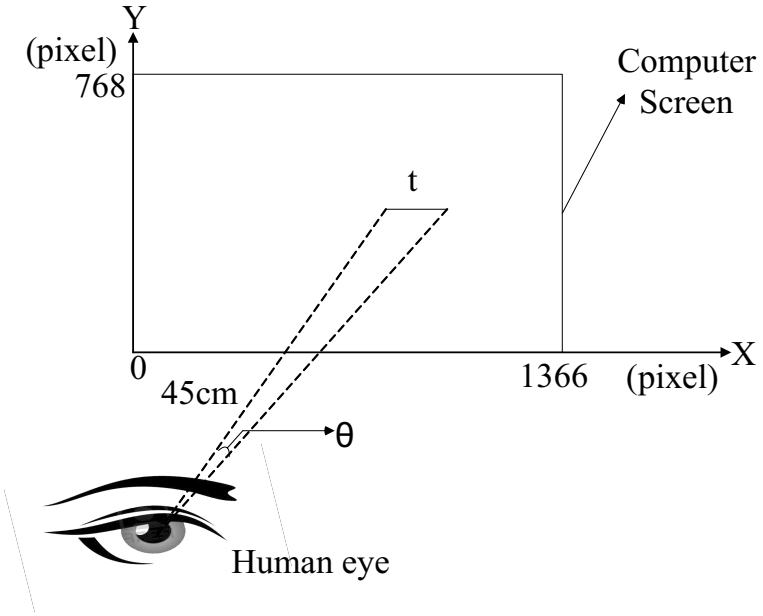


Figure 4.9: Calculation Method of Degree Error

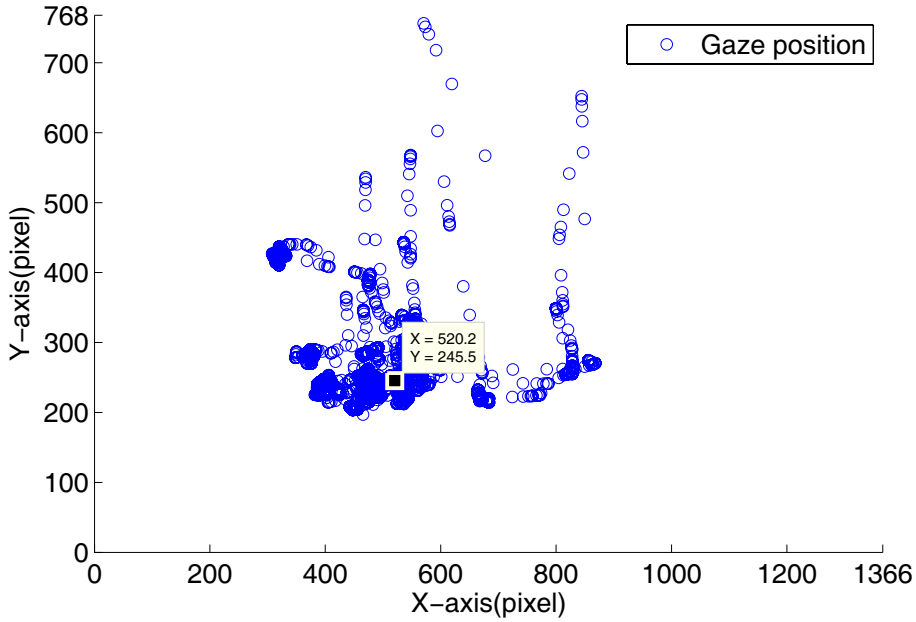


Figure 4.10: Distribution of User's Gaze Position

saliency map got based on image processing (explained in the next chapter), user's visual attention regions are decided. An example of gaze distribution is show in Fig.4.10, where the axis X and Y is 0 to 1366 and 0 to 768 respectively, corresponding with the screen resolution. In Fig.4.10, we can see at the users intention region, whose center is marked in the figure at (520.2, 245.5), the points distribution density is high. Note that the sampling

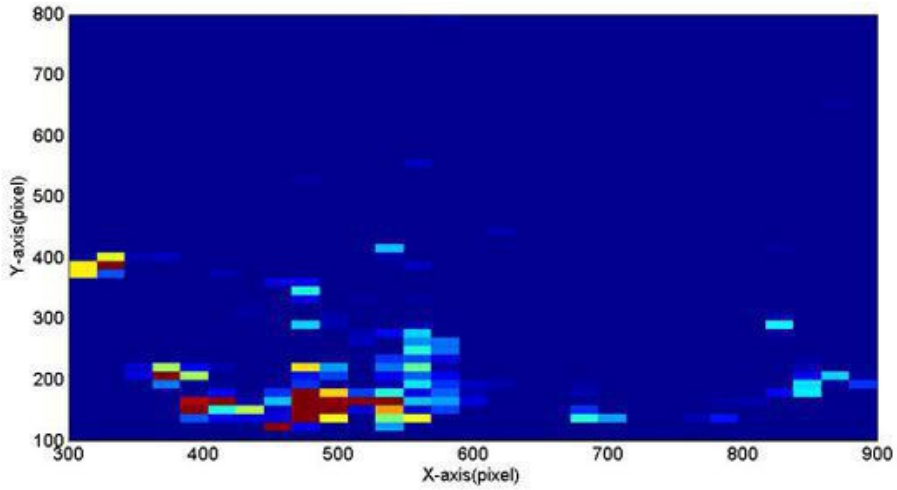


Figure 4.11: Statistics of Distribution of User's Gaze Position

Table 4.4: Specifications of Remote Eye Tracking Device

QG-PLUS Mini (DITECT)	
Dimensions:	29×2×2.5 cm
Typical operating distance:	65 cm
Accuracy:	0.5° cm
Method:	Dark pupil, binocular tracking
Frame rate:	80 FPS

time period is 4s. To observe the density of points intuitively, we divide the screen into some small regions by an optimal grid, which is 49×28, and make a statistical density map of the number of points in each grid. Fig.4.11 is the gaze distribution after normalization.

4.2 Remote Camera Based Eye Tracking

The remote camera based eye tracking system only consists by one unit. With the simplicity of a single USB interface, it offers valuable metrics such as gaze position, stagnation times, blink rates, pupil size.

4.2.1 Remote Eye Tracking Device

Eye tracking device used in this research is shown in Fig.4.12. As we can see in Fig.4.12, there are infrared LED array on both sides of the device, which in order to provide reference

Figure 4.12: Remote Eye Tracking Device: QG-Plus Mini (produced by DITECT)

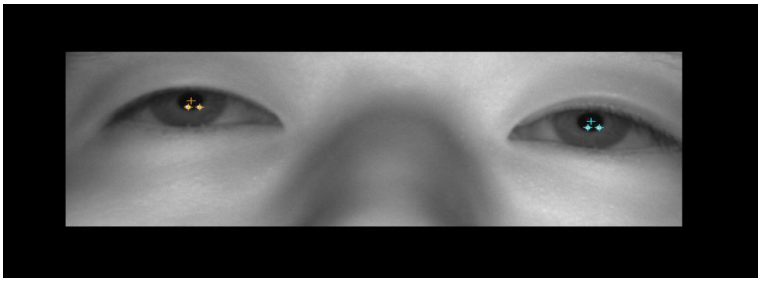


Figure 4.13: Position Detection of Two Pupils

positions for pupil detection. Eye gaze will be recorded as 2-dimensional Cartesian coordinates on the screen at an average rate of 70 data points per second. The number of data points here is set to 70 in order to get the best combination of head movement tolerance and fast tracking speeds. Detailed specifications of the device are shown in Table.4.4

It is worth to note that the system can compensate a small range of head movement by measuring the angle of line which decided by two eyes position and improve the precision of the tracking results. The eye pupil detection is shown in Fig.4.13. Take right eye for example, the cross on the upper is the center of pupil. And the two crosses on the lower position are reference points provided by infrared LED arrays.

4.2.2 Analysis of Gaze Distribution

In order to analyze the relationship between user's eye gaze and intention, we analyze the gaze distribution of user when looking at images. Experiments are conducted by using the

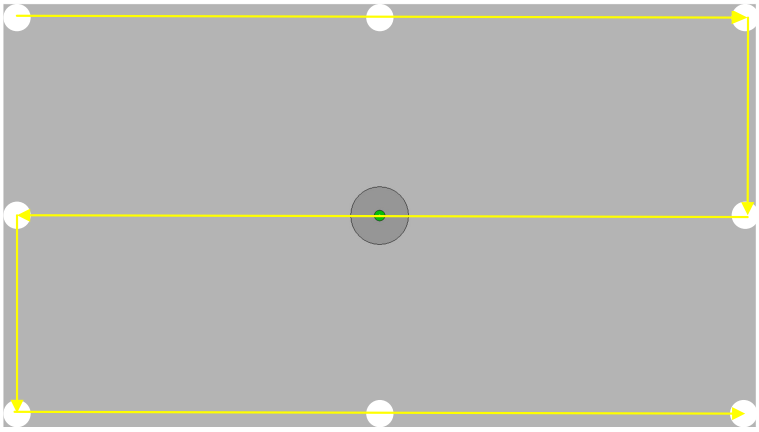


Figure 4.14: Scene of Calibration Process

eye tracking device mentioned in section 4.2.1. Computer used for processing is a notebook PC with Intel Core i3-380M CPU, 2 GB RAM and Microsoft Windows 7 operating system. The distance between user and computer screen is 60cm. As shown in Fig.4.12, the system works by placing the device in front of user and under the computer screen. Subject for following typical experiments is a male who is 31 years old without glasses.

Just as the wearable device based mode, the calibration process is also needed for this eye tracking system. Reference points can be set to 5, 9 or 16 points in this system. Of course, more reference points set, higher accuracy will be got, but more calibration time taken at the same time. The calibration process is shown in Fig.4.14.

After calibration, we do the experiment by asking participant looking at example images and record gaze position data at the same time. Then based on the gaze position data of participate, we analyze the gaze distribution in three methods.

(1)Order analysis

Address to where and how long the participant's gaze staying and the order of looking in the experiment process, gaze analysis are realized by counting, digitizing, visualizing of the gaze position data. The results of gaze analysis can also illustrate that where and what kind of order the participant performed in experiment. Two examples of experimental results are shown in Fig.4.15.(a). And in Fig.4.15.(a), the range of region considering as same stay position is 50 pixels. And the regions where stay time less than 0.5s are ignored

(a) Gaze Analysis

(b) Stagnation Map

(c) Area Analysis

Figure 4.15: Two Examples of Gaze Distribution Analysis Results when Scanning Image

and not reflected in the result. We can see that the numbers and lines illustrate the order for the looking while size of the points stand for the stagnation times.

(2)Stagnation map

Stagnation map is calculated in order to make it easy to understand that which part the participant has been focusing on in the experiment. Two examples experimental results

of stagnation map also shown in Fig.4.15.(b), and the range of region considering as same stay position is 10 pixels while the threshold of ignoring time is 0.05s. The darkness of the color indicates the ranking: the darkest red marks the highest value, the green stand for lower.

(3)Area analysis

The participant paid how much attention and saw how many times in an area are also an important issue for the gaze distribution analysis. Therefore, we also do area analysis after experiment to visualize them. The experimental results are shown in Fig.4.15.(c). There are two different method for area analysis. For the image with complicated structure or pattern, the image is dividing into several regions in equal size. The staying time of gaze and times of scanning in each region are counted and analyzed, as shown on the left side of the figure. And for the image with simple structure or obvious object, analysis only made in some specific regions as show on the right side of the figure.

4.3 Conclusions

In this chapter, two different modes of eye tracking system are described. For the wearable device based eye tracking mode, a low cost eye tracking system is designed by us. We also use neural network to replace the traditional spatial mapping method in the calibration process. In the experiment we confirm that after 1000 trials learning by using enough reference points, a better calibration result can be obtained.

In addition, in order to illustrate the relationship between gaze movement and attention, the gaze distribution when participant looking at an image is also analyzed by using both two systems. According to the experimental results in this chapter, we find that gaze coordinates output of eye tracking system strongly indicate participate's focus of attention. However, it is difficult to interpret eye movement patterns. In other words, we can not understand participate's attention only by using real time gaze position. Just for this reason, the feature of eye movement in a period is needed.

Chapter 5

Proposed Intention Recognition System

5.1 Introduction

Intention recognition is a task of recognizing the intentions of a human or an agent. The task usually achieved by analyzing some or all the actions or changes of some features. The basic structure of an intention recognition system is usually composed of three parts, as shown in Fig.5.1.

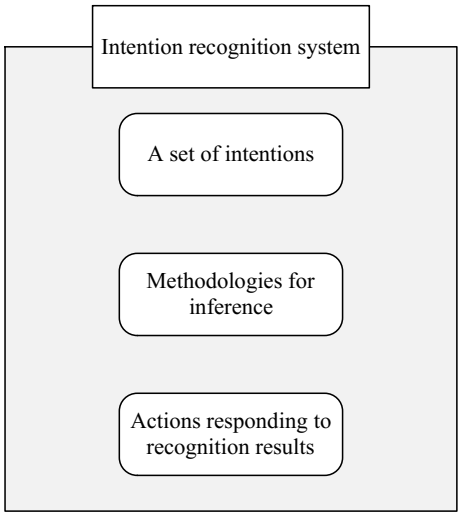


Figure 5.1: Structure of Intention Recognition System

5.2 Intention Recognition with Eye Tracking and Object Recognition

Specifically, intention recognition can be realized by two steps: founding a set of features or actions and inferring based on a theory. In this section, for the first step, we find the initial set of intentions by object recognition in the image of robot vision and choose human's gaze as actions. For the second step, we infer human intention based on fuzzy theory.

5.2.1 Object Recognition

Before the intention processing, the initial set of possible intentions must be founded. The set also should depend on the situation at hand. In this research we create the set by recognizing objects in image showing to user. Usually, objects can be divided into two types, known objects and unknown ones. To obtain the intention set, one or some known objects are organized, which also means that the objects have been learning by the system.

After learning object recognition[66], when an object appeared in the view field of

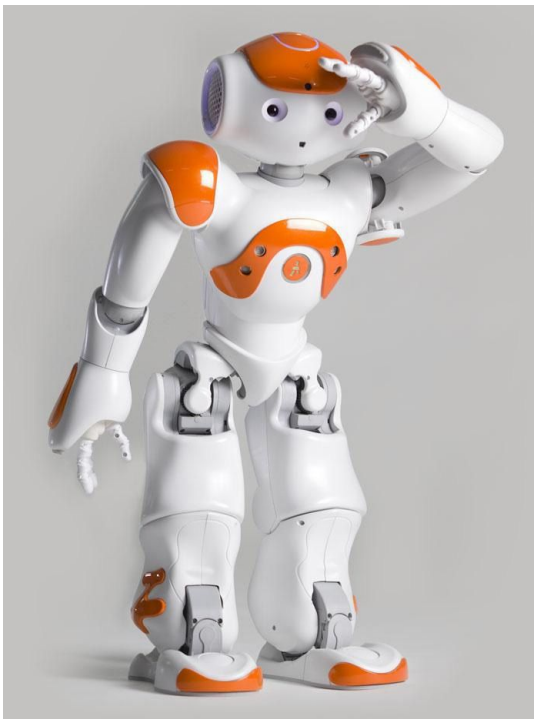


Figure 5.2: Humanoid Robot: NAO

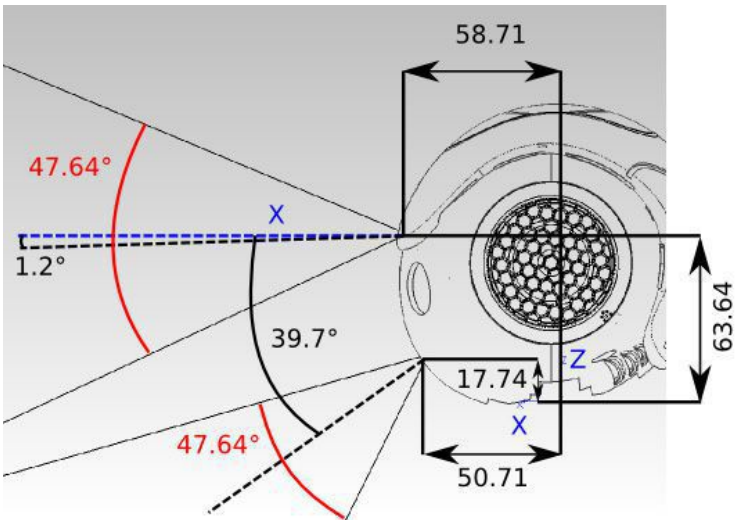


Figure 5.3: Field View of NAO[66]

NAO, we can get the position of object in NAO's coordinate system. The feature points of object's contour can also get at the same time. This process is achieved by using the ALVisionRecognition module from NAO SDK, which is a vision module in which NAO tries to recognize different pictures, objects sides or even locations learned previously. This module is based on the recognition of visual key points and is only intended to recognize specific objects that have been learned. Examples of object recognition are shown in Fig.5.4. In this figure, there are two balls on the table, pink one and yellow one. When balls appeared in view field, the rectangle will figure out the region and position information.

5.2.2 Intention Recognition by Fuzzy Inference

After object recognition, the regions where objects exist are considered as the initial set of intention regions where user may pay attention and ready for using by intention inference.

By using the eye tracking device mentioned in last chapter, we analyze the data of user's gaze position on the computer screen in a small period. In fact, the user's gaze will have a longer time staying and higher frequency appearing in the intention region than in not noticed region. Based on this, we use the frequency and continuous staying time as the factors for intention inference.

So far, the researches on intention recognition can be mainly divided into three classes

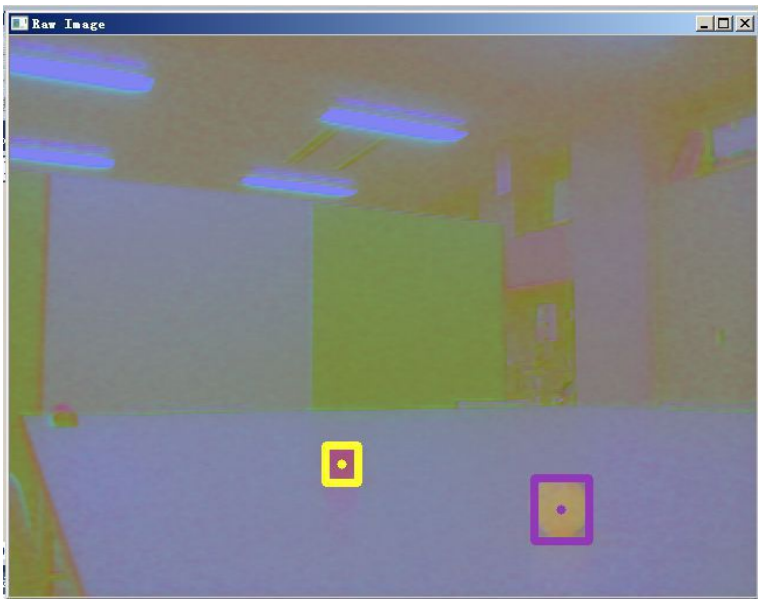


Figure 5.4: Example of Object Recognition

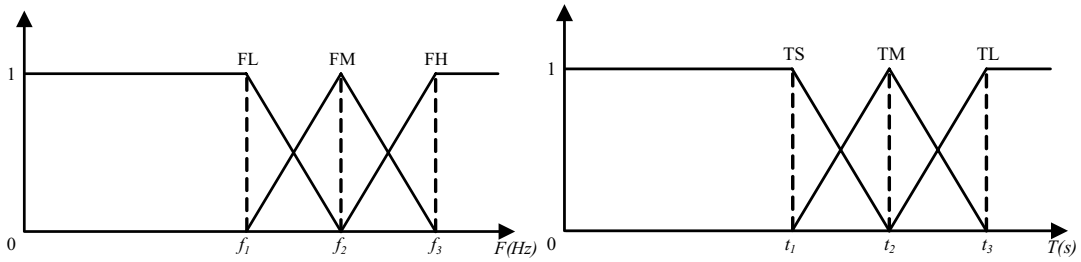
according to the second component in Fig.5.1, which are logic-based, case-based and probabilistic approaches[67].

According to the uncertainty of human's conceptual judgment and reasoning way of thinking, by using fuzzy sets and fuzzy rules in reasoning and making a fuzzy comprehensive judgment, we can solve the complicated problems which are difficult for normal methods such as intention recognition. Thus, in this thesis, we apply the fuzzy inference as the method to recognize human's intention. Fuzzy inference is based on fuzzy logic and resembles human reasoning in its use of approximate information and uncertainty to generate decisions[68].

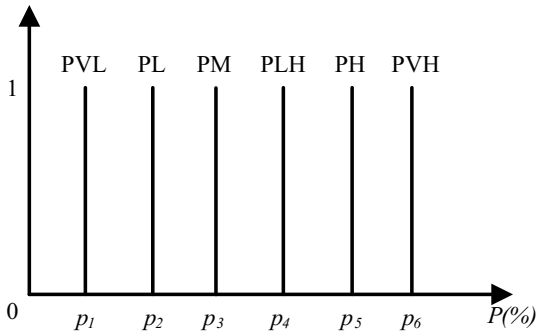
Fuzzy rules are used to describe the relationship between user's gaze and his/her intention. The fuzzy rule map is shown in Table 5.1. Fig.5.5 shows the membership functions and singletons. We use the gaze appearing frequency F and the gaze continuously staying time T in the regions got by object recognition part as the input of fuzzy rules. And the output is the possibility of current region is the intention one.

Table 5.1: Fuzzy Rule Map for Intention Recognition

T \ F	FL	FM	FH
TS	PVL	PL	PLH
TM	PL	PM	PH
TL	PLH	PH	PVH



(a) Membership Functions in IF Part



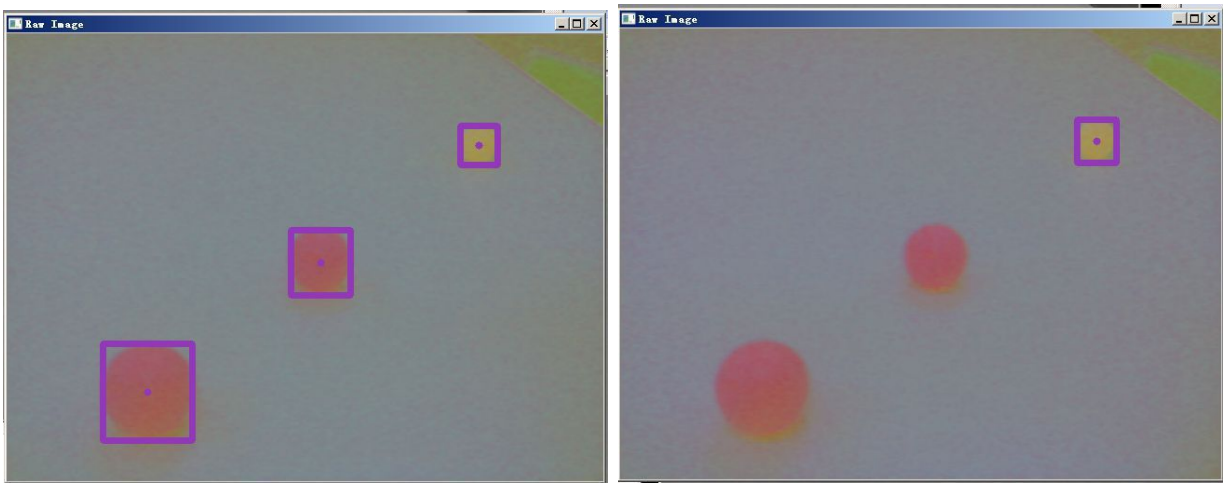
(b) Singletons in THEN Part

Figure 5.5: Fuzzy Sets for Intention Recognition

5.2.3 Experimental Results

In the experiment, after a calibration process of eye tracking, we ask subject to look at the scene image from NAO's camera which is shown on computer screen and control NAO's head rotate by gaze. When an object or some objects are found, the gaze frequency and continuous time in the objects existing regions in a certain period are used as the input of the fuzzy inference system which mentioned above.

Then after an intention recognition result is given, the robot will get the object by using its hands. It is worth to note that there are three types of coordinate systems for NAO; FRAME-TORSO, FRAME-WORLD and FRAME-ROBOT. When creating a command



(a) Object Recognition without Fuzzy Inference

(b) Intention Recognition with Fuzzy Inference

Figure 5.6: Intention Recognition by Fuzzy Inference

for NAO, much attention needs to be placed on the space used to define the command. In this process the position of object gotten from NAO is according to FRAME-TORSO, also the same for operating hands to grasp an object. Here FRAME-TORSO is one of the 3 spatial references used by NAO's motion components.

Fig.5.6.(a) shows the various object recognition results in our experiment. We can see that there are three objects have been recognized before intention inference. And Fig.5.7 shows the distribution of user's gaze positions when looking at this scene. The output result of intention recognition by fuzzy inference is shown in Fig.5.6.(b). According to Fig.5.6.(b), user's intention is at the region where marked by a rectangle. According to Fig.5.6.(b), user intended on the yellow ball in this experiment. It is worth to note that the result in this experiment is a typical one. It also means that experimental result may changed according to different subjects.

After intention recognition, an experiment of getting a ball for user is also conducted based on intention recognition result by fuzzy inference showed in Fig.5.6. The experiment scenes are shown in Fig.5.8.

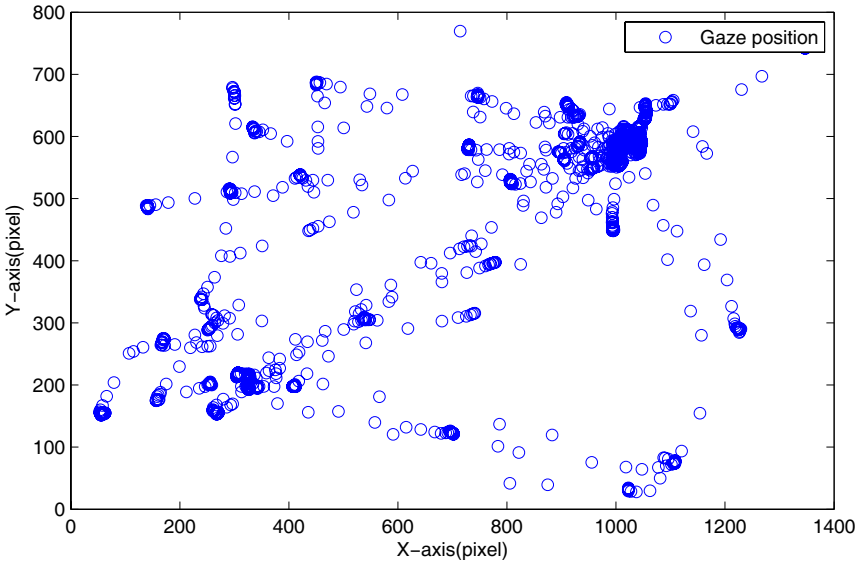


Figure 5.7: Distribution of User's Gaze Position

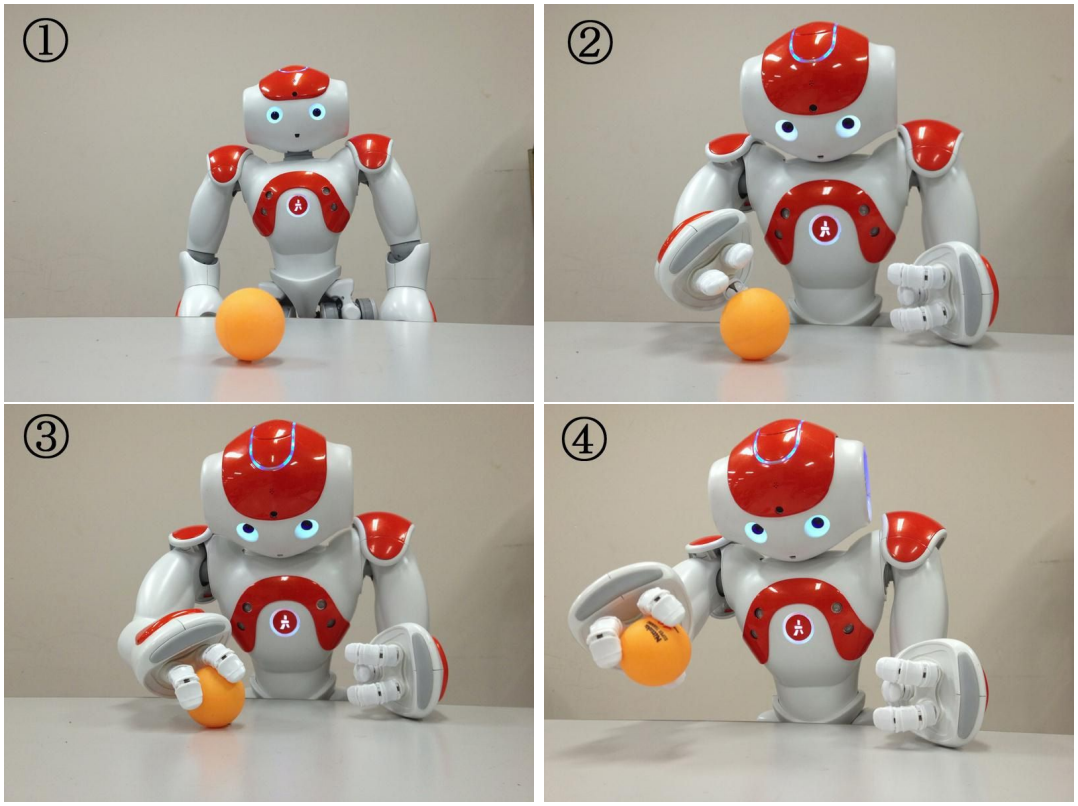


Figure 5.8: Experiment Scenes of Getting Object by NAO

5.2.4 Discussion

In this section, we proposed a method of intention recognition by using eye tracking and object recognition. As shown in the experimental results, participant's intention can be recognized through this method. But there still several problems exist. The first one is that the robot can not do the corresponding task according to the recognition result directly until it confirmed by participate. This is because the possibility of error also exists. Another problem is that object recognition process requires object learning in the prior period, which limits its widespread application. Therefore, how to remove the restrictions on including non subjective factors becomes an important issue for improving the system.

5.3 Attention Prediction with Saliency Map

Most attention models are based on a saliency map and a dynamical process for visiting saliency maxima. Itti et al. (1998)[26] introduced a model for bottom-up selective attention based on serially scanning a saliency map, which is computed from local feature contrasts, for salient locations in the order of decreasing saliency. The saliency map is entirely based on features of image and was originally designed to explain converting attention on simple stimuli. A lot of researches on saliency map are getting some features of image and combining them by simply sum in mathematics[23].

But this method also has its weakness. For example, the saliency map of an image is based on the three features which are color, intensity and orientation. The method of simple sum of them gives them the same importance at the same time. But based on experiments of the paper [69], most people do not pay equal attention to all of them. Actually, for example, when the color feature in an image takes more attention from observers than other two features, the method may not be very reasonable. In order to solve this problem, we propose two methods to compute saliency map by using fuzzy inference and FNN based on the features of image. In this way, the importance of all features can be reflected in fuzzy rules or weights.

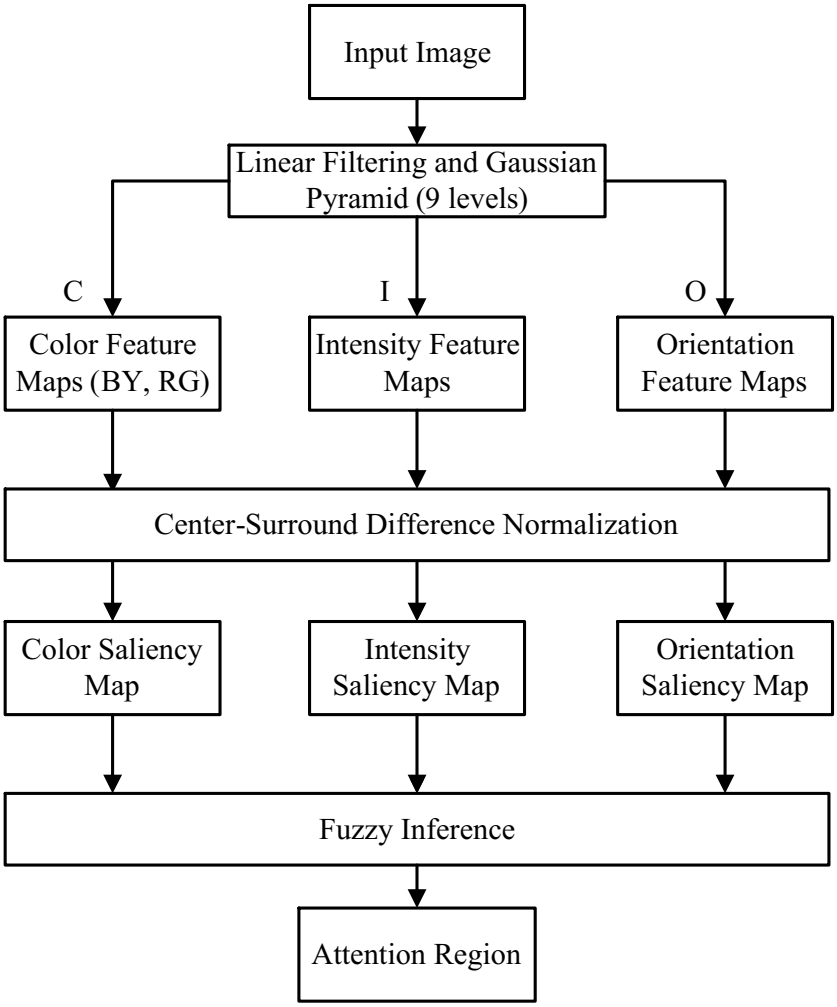


Figure 5.9: Architecture of Computing Feature Maps

5.3.1 Feature Maps

This section presents the details of our computational framework for building feature saliency maps. For a color input image, we compute feature maps for color, intensity, and orientation contrasts at different scales, as shown in Fig.5.9.

In Fig.5.9, the linear filter is used in order to compute center-surround differences of various features at 9 scales. In this paper, the input image is sub-sampled into a dyadic Gaussian pyramid by convolution with a linearly separable Gaussian filter and extraction by a factor of two[19], which also means that only half of the image pixels are sampled and processing speed is reduced by half. Gaussian pyramids are used in order to compute center-surround differences of various features at various scales. The conventional way

of creating the levels of the pyramid consists of two separate steps, convolution with a separable Gaussian filter followed by decimation.

Suppose the image is represented initially by the array g_0, which contains C columns and R rows of pixels. Each pixel represents the light intensity at the corresponding image point by an integer I between 0 and K. This image becomes the bottom or zero level of the Gaussian pyramid. Pyramid level 1 contains image g_1, which is a reduced or low-pass filtered version of g_0. Each value within level 1 is computed as a weighted average of values in level 0 within a 5-by-5 window. Each value within level 2, representing g_2, is then obtained from values within level 1 by applying the same pattern of weights. And so on, 9 levels are obtained.

After filtering, the three features of an image have their values at each position according to the input image, which are divided into 9 levels of pyramid ready to be calculated. Then, in this paper, the color feature is reflected by two values defined by us, which are red-green and blue-yellow opponencies. If r, g, b and y are the red, green, blue and yellow values of the input color image respectively. Then the color map of one level can be calculated according to the following equations:

$$M_{r-g} = \frac{r-g}{max(r,g,b)}, \tag{5.1}$$

$$M_{b-y} = \frac{b-min(r,g)}{max(r,g,b)}, \tag{5.2}$$

where M_{r-g}, M_{b-y} stand for red-green and blue-yellow opponencies. And *min(r,g)* is used to reflect the information of yellow color because yellow is perceived as the overlap of red and green in equal parts, so that the amount of yellow contained in an RGB pixel is given by *min(r,g)*. And note that the definitions deviate from the original model by Itti[26].

The intensity map of one level is calculated as:

$$M_i = \frac{r+g+b}{3} \tag{5.3}$$

These operations are repeated for each level of the input pyramid to obtain an intensity pyramid with also 9 levels.

Local orientation maps M_o is obtained by applying steerable filters to the intensity pyramid levels M_i[67].

After getting M_{r-g}, M_{b-y}, M_i and M_o, in order to yield the feature maps, we simulate the center-surround receptive fields by subtraction between two maps at the center (c) and the surround (s) levels in these pyramids. They can be calculated as:

$$\begin{aligned} F_{l,c,s} &= N(|M_l(c) - M_l(s)|), \\ l \in L &= L_C \cup L_I \cup L_O, \end{aligned} \tag{5.4}$$

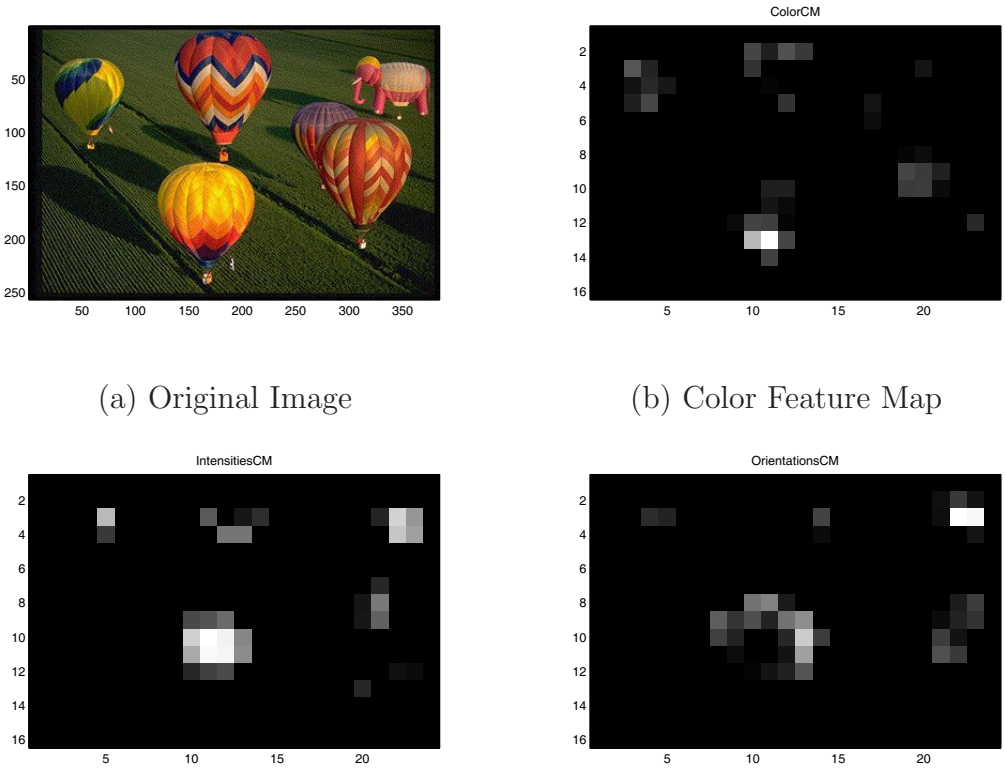
where

$$\begin{aligned} L_C &= \{I\}, \\ L_I &= \{r-g, b-y\}, \\ L_O &= \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}. \end{aligned} \tag{5.5}$$

Note that N is an iterative, nonlinear normalization operator, simulating local competition between neighboring salient locations. Each iteration step consists of self-excitation and neighbor-induced inhibition, implemented by convolution with a difference of Gaussians filter, followed by rectification. For the simulations in this paper, between one and five iterations are used.

Finally, by summing over the center-surround combinations and normalizing again according the results obtained in Eq.(5.4), the feature maps of color, intensity and orientation can be obtained according to Eq.(5.6) as C_c, C_i, C_o, respectively. Center-surround receptive fields are simulated by subtraction between two maps at the center (c) and the surround (s) levels in these pyramids, as shown in Eq.(5.4). In Eq.(5.6), for the general features color and orientation, the contributions of the sub-features are summed and normalized once more to yield conspicuity maps. For intensity, the conspicuity map is the same as in Eq.(5.4). Fig.5.10 shows an example of three feature maps mentioned above.

$$C_i = F_i,$$



(a) Original Image

(b) Color Feature Map

(c) Intensity Feature Map

(d) Orientation Feature Map

Figure 5.10: Example of Feature Maps Obtained from Image

$$\begin{aligned} C_c &= N(\sum_{l \in L_c} F_c), \\ C_o &= N(\sum_{l \in L_o} F_o). \end{aligned} \tag{5.6}$$

5.3.2 Attention Prediction with Saliency Map by Fuzzy Inference

In order to reflect the importance of image features, we propose fuzzy inference method for saliency map computing at first. In this way, the importance of image features can be reflected in fuzzy rules.

(1)Fuzzy Inference

Fuzzy inference for Saliency Map is based on fuzzy logic and resembles human reasoning in its use of approximate information and uncertainty to generate decisions. In the beginning

of section 5.3, we have pointed out the weakness of ordinary combination method of feature maps. It has not an important distinction between various features, especially when a feature is more important comparing with others.

In the building stage of feature maps, it has little sense to use fuzzy theory as a classifier, but in the combination stage, using fuzzy theory to infer visual attention region can highlight the significance of various features of image and obtain better results. The greatest difference between mathematical sum method and fuzzy inference method is that the importance reflected in every feature map is different. These are tuned by fuzzy rule according to feature values. In the fuzzy inference method, the importance of color feature map is higher than intensity and orientation ones, to avoid getting bad result when the situation that color feature is low while other two are high.

In this research, We use the feature variables from color feature map (C_c), intensity feature map (C_i) and orientation feature map (C_o) in the *IF* part while the output value in *THEN* part is value of region saliency map (S_m)[70]. Each value of region saliency map is decided by fuzzy rule as shown in Table 5.2 and Fig.5.11. In Fig.5.11, in membership functions in IF part, the values of l, m and i are used to divide the input space into three fuzzy subsets and assigned linguistic terms to them. We obtained the values based on expert's experience. We set values based on the analysis results of the data obtained in the training process. Take l for example, its values are as following.

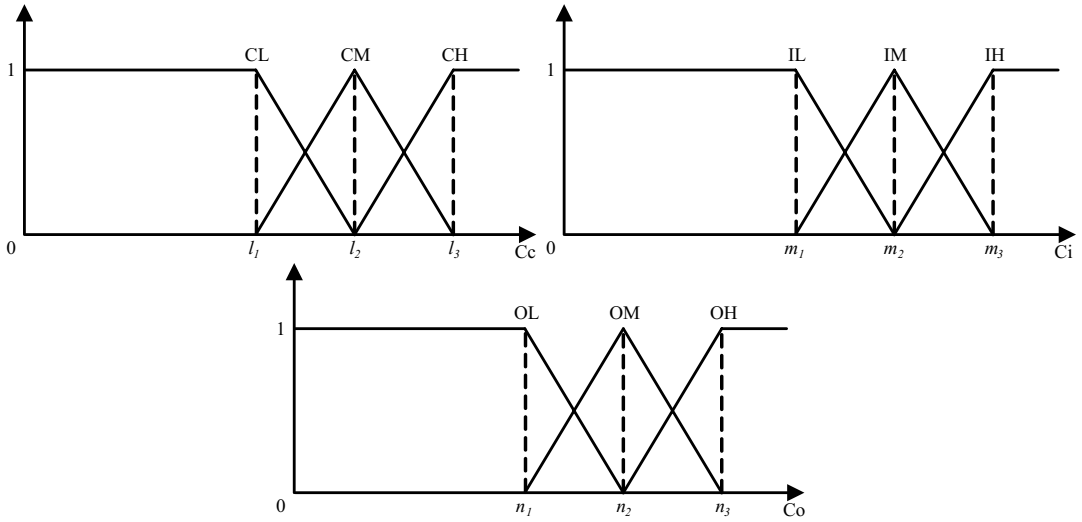
$$\begin{aligned} l_1 &= 0.4(l_{max} - l_{min}), \\ l_2 &= 0.6(l_{max} - l_{min}), \\ l_3 &= 0.8(l_{max} - l_{min}). \end{aligned} \tag{5.7}$$

where (l_{max}) and (l_{min}) are maximum and minimum value of color feature value respectively.

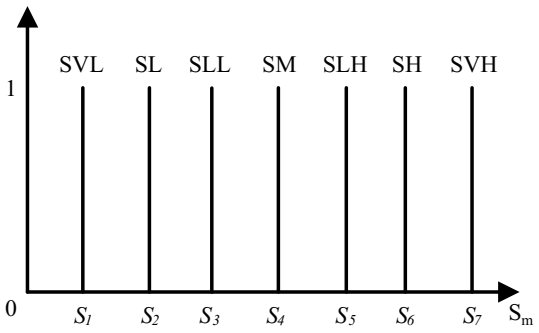
The fuzzy rules should reflect the importance of the key feature. In addition because the fuzzy rules cannot change in the inference process, in this research we obtain the fuzzy rules based on expert's experience. The fuzzy rules are generated from training input-output pairs.

Table 5.2: Fuzzy Rule for Saliency Map Inference

C	O / I	OL	OM	OH
CL	IL	SVL	SL	SLL
	IM	SL	SLL	SLL
	IH	SLL	SLL	SM
CM	IL	SL	SLL	SM
	IM	SLL	SM	SM
	IH	SM	SM	SLH
CH	IL	SM	SLH	SH
	IM	SLH	SH	SH
	IH	SH	SH	SVH



(a) Membership Functions in IF Part



(b) Singletons in THEN Part

Figure 5.11: Fuzzy Sets for Fuzzy Inference of Saliency Map

We proceed in the following manner. First, we get input and output data by doing lots of experiments. In the experiments, we use McGill calibrated color Image Database[71] as our image database. The images in this database all have been calibrated. And by using the method mentioned in section 5.3.1, we get the color, intensity and orientation feature values of each image. Then divide each input space into three fuzzy subsets and assigned linguistic terms to them. According the saliency map of each image, the values of each pixel make up the output space. Output space is divided into seven subsets as SVL, SL, SLL, SM, SLH, SH and SVH, which means very low, low, little low, medium, little high, high and very high. Second, we show the images to subjects and ask the region they intended in the experiment process. Actually, they are asked to give three regions in order. Then saliency value of pixel in user's intention region is raised according to region's order. At last, by analyzing the data obtained in previous steps, we generate fuzzy rules using the linguistic terms assigned. The sample rules are as following:

IF C is low AND I is low AND O is low THEN S is very low;

IF C is medium AND I is low AND O is high THEN S is medium;

IF C is high AND I is medium AND O is low THEN S is little high;

- Output Fuzzy Labels of Region Saliency Map(S_m)

  SVL : Value of Region Saliency Map Very Small
  SL : Value of Region Saliency Map Small
  SLL : Value of Region Saliency Map Little Small
  SM : Value of Region Saliency Map Medium
  SLH: Value of Region Saliency Map Little Large
  SH : Value of Region Saliency Map Large
  SVH : Value of Region Saliency Map Very Large

(2)Experimental Results

We conduct several experiments to demonstrate the inference result of the proposed method and also compare the performance of the proposed method with Itti's model[23]. As mentioned in last section, we get the various feature maps at first. Here, four different image examples are used as the input. The input image is processed for low-level features at multiple scales, and center-surround differences are computed according to Eq.(5.4). Then,

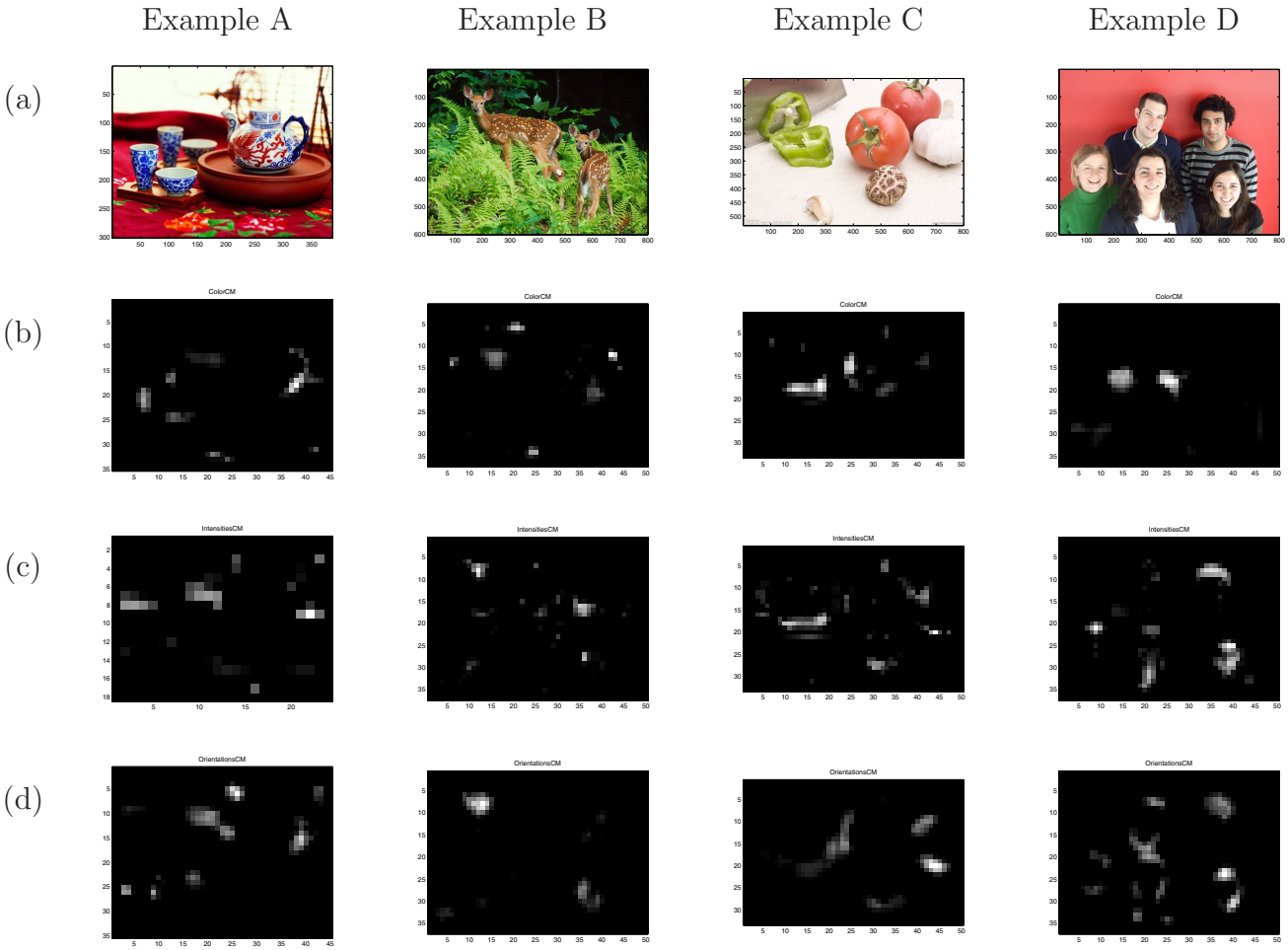


Figure 5.12: Four Examples of Feature Maps: (a)original image, (b)color feature map, (c)intensity feature map, (d)orientation feature map

the resulting feature maps are combined into feature saliency maps according to Eq.(5.6), which is shown in Fig.5.12.

After getting the feature saliency maps, the region located in the saliency map is selected for the highest saliency value by proposed fuzzy inference. After segmentation around the most salient region location, this saliency map is used for obtaining a smooth object mask at image resolution and for showing to participate. The parameters of fuzzy rules (l_i, m_i, n_i) are chosen based on the average value of each feature map. The results of saliency map and attention region by summing feature maps method and fuzzy inference method are shown in Fig.5.13. The attention regions are marked by yellow lines while red lines express

Figure 5.13: Four Examples of Saliency Maps and Attention Region: (a) original image, (b)saliency map by sum feature maps method, (c)saliency map by fuzzy inference, (d)attention region by sum feature maps method, (e)attention region by fuzzy inference

the order of them.

As we can see in the Fig.5.13, there are only little differences between the saliency maps of two methods. And for examples A and B, the approximate locations of attention regions and the orders are basically the same between two methods while the little difference is the shape and size of region. This is because the color, intensity and orientation feature are

all reflect obviously in regions marked of these two images compared with the rest regions, which also means that the differences of importance for the three features are small. So the method we proposed has not functioned very efficiently. But in the result of example C we can see that the first attention region decided by fuzzy inference method is nearby the tomato while the one decided by sum method is nearby the garlic. This is illustrated that our method is worked because the region nearby the tomato is more conspicuous at the feature at color.

The result of example D shows that the attention regions of both two methods have obvious differences while the order is different also. But only from these results we cannot recognize yet whether our proposed method is better than Itti's or not. In other words, we cannot assert that the proposed method has higher accuracy for the computation of saliency map than the conventional one yet. So we conduct another experiment to verify it.

In the following experiment we ask 5 males, who are between 20 to 30 years old, to look over all of the 50 images while using the eye tracking device. When they looking at images, the gaze positions of them at different time are calculated and recorded for distribution analysis at last. After all the experiments they will be showed the results of attention region got by the two methods and asked to compare with the ones they actually attending and looking at in the experiment process. Two factors are rated in experiments. One is visual attention region, which in order to illustrate whether saliency region predicted by fuzzy inference is in accordance with the actual contour of object. The other one is order of regions which illustrate the order of user attention movement. Finally, an evaluation of user's attitude to the results is carried on and shown in Fig.5.14. Every factor of them has five ranks and represented by 1∼5 from worst to best in these figures. In Fig.5.13, the results (d) and (e) give the attention regions predicted by two different calculation methods. And in following experiment, subject is asked to look at the same images used in Fig.5.13. At the same time, subject's gaze positions are recorded for further analyzing. Furthermore, the gaze distributions of participant also be analyzed as shown in Fig.5.15.

We can see from the evaluation results and that the performance of our proposed method is better than the traditional method. This is also illustrated that the proposed fuzzy

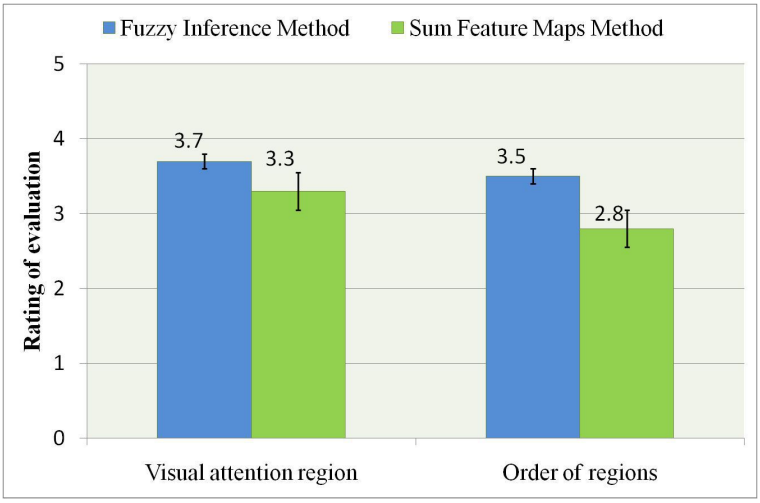


Figure 5.14: Standard Deviation for Evaluation Results of Attention Regions Obtained by Two Methods

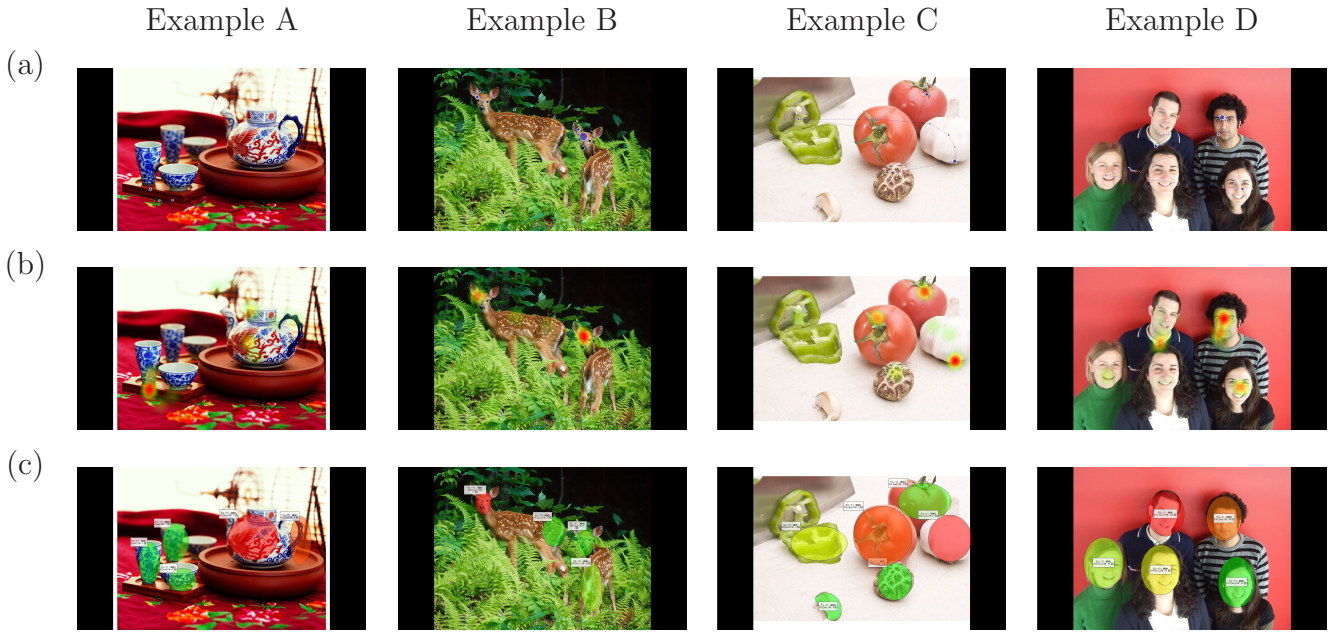
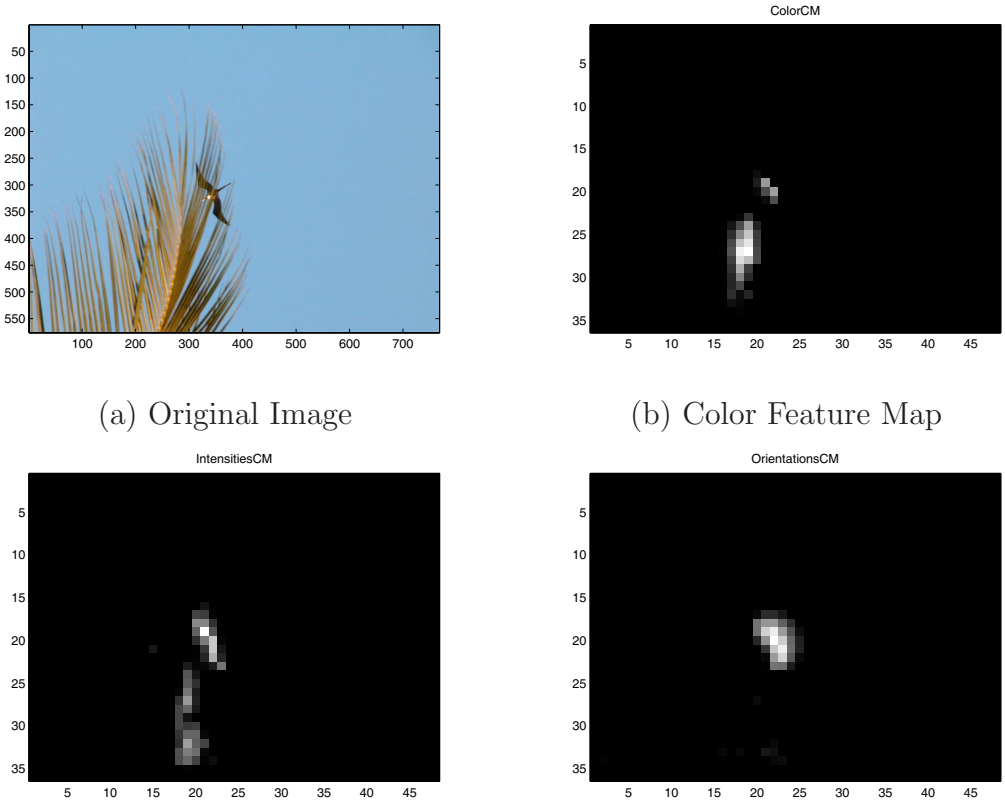


Figure 5.15: One Example of Gaze Distribution Analysis Results: (a)oder analysis, (b)stagnation map, (c)area analysis

inference method can improve the performance of attention region prediction at some aspect.



(a) Original Image

(b) Color Feature Map

(c) Intensity Feature Map

(d) Orientation Feature Map

Figure 5.16: Specific Example of Feature Maps

5.3.3 Attention Prediction with Saliency Map by FNN

According to the method proposed in section 5.3.2, the importance of color feature can be reflected well. But the method is only suitable for specific images, which means is not universal. For example, when the color feature values are lower compared with the other two, the method mentioned above will decide the saliency values according to the higher two features. In general cases they will have correct results, but when special cases such as the relative difference value is larger than other ones' although has lower feather values itself. As we can see in Fig.5.16, the color and orientation feature values are higher than intensity one's in this case, but according to the image and the gaze distribution result shown in Fig.5.17, we know that the attention object of human should be the bird in the image, which also means that the saliency region should be decided by the intensity feature.

Figure 5.17: Analysis of Gaze Distribution with Specific Example Image

Therefore, we propose another method by using FNN to solve this problem. In this way, the importance of all features can be reflected in fuzzy rule with the human decision making model by the conceptual framework of fuzzy logic. The overall procedural flow of proposed approach is summarized in Fig.5.9. Actually, comparing with the method proposed in section 5.3.2, the only difference is FNN is used to instead fuzzy inference. The saliency map in this section is generated by a trained FNN using the three saliency maps as inputs. The training method of FNN will be explaining in the following section.

(1)Fuzzy Neural Network

It has been shown that a fuzzy system can approximate any continuous real function defined on a compact domain by covering its function graph in input-output space using a set of if-then fuzzy rules. Theoretically, these fuzzy rules can always be discovered, but in practice we may have no idea on how to initialize these rules. Thus, it is crucial to have an adaptive fuzzy system which can produce the required rules automatically[72, 73]. A FNN system is a learning machine that finds the parameters of fuzzy rules by exploiting approximation techniques from neural networks.

In the feature maps building stage, it has little sense to use fuzzy theory as a classifier, but in the combination stage, using FNN to infer visual attention region can lead to the

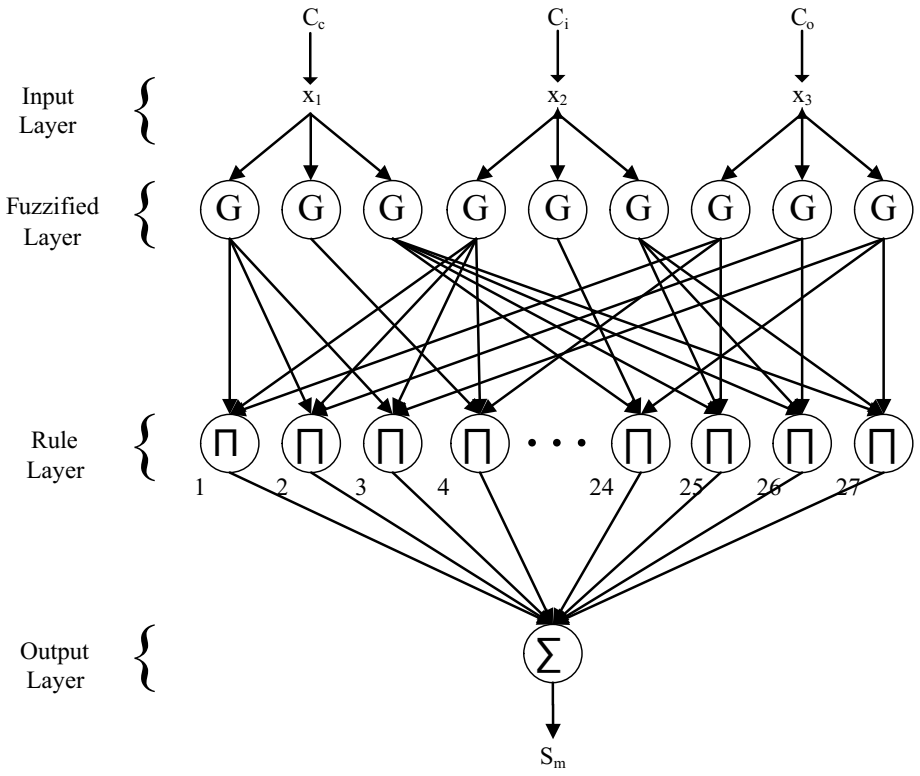


Figure 5.18: The Structure of the FNN

results which can reflect the importance of each feather map in the saliency feature that images have. The greatest difference between mathematical sum method and FNN method is that the importance reflected in every feature map is different. These are tuned by fuzzy rules and connection weights of the network according to feature values. In FNN method, the importance of each feature map is different according to different images.

We also use the feature variables from color feature map (C_c), intensity feature map (C_i) and orientation feature map (C_o) as input while the output is a value of region saliency map (S_m). Every value of region saliency map is decided by FNN as shown in Fig.5.18 where G stands for Gaussian Function[74, 75].

As shown in Fig.5.18, the FNN structure with three input variables, three input nodes of fuzzified layer for each input variable, 27 rule nodes of hidden layer, and one output node of output layer. A typical format for a fuzzy rule base consists of a collection of fuzzy

IF-THEN rules in the following form:

$$IF\, x_1 is A^j_{111}, \cdots, and x_n is A^j_{nml}, THEN\, S^j_m is \beta^j \tag{5.8}$$

where A^j_{nml} and β^j are fuzzy sets and x_i and S^j_m are the input and output of the fuzzy inference rule, respectively. And n, m, l stand for the node number in fuzzified layer of (C_c), (C_i) and (C_o) while i is the input node number for rule layer and j is the output node number of rule layer.

(A)Fuzzified Layer

This layer uses a Gaussian function as a membership function, so the output of the ith term node associated with x_i is:

$$\mu_{A_{ijk}} = exp(-(\frac{x_i - m_{ijk}}{\sigma_{ijk}})^2) \tag{5.9}$$

where m_{ijk} and σ_{ijk} denote the mean (center) and variance (width) of A_{ijk}, respectively. And i, j, k have the similar meaning as n, m, l in Eq.(5.8).

(B)Rule Layer

This layer implements the links relating preconditions (fuzzified layer) to consequences (output layer). The connection criterion is that each rule node has only one antecedent link from a fuzzified node of a linguistic variable. Hence there are 27 rule nodes in the initial form of FNN structure. We mention that there is still no weight adjustment in this layer. The output of the jth rule node is:

$$out^3_j = \prod_{i=1}^{n} \mu_{A_{ikl}}(x_i) \tag{5.10}$$

where the superscript 3 of out stands for the input number is 3, and i, k, l have the similar meaning as n, m, l in Eq.(5.8). Only noted that l is determined by the connection criterion.

(C)Output Layer

All consequence links are fully connected to the output nodes and interpreted directly as the strength of the output action. This layer performs defuzzification to obtain the numerical output:

$$S_m = \sum_{j=1}^{m} \beta^j \prod_{i=1}^{n} \mu_{A_{ijk}}(x_i) \tag{5.11}$$

where m is the number of fuzzy IF-THEN rules and n is inputs number.

(2)Supervised Learning of Fuzzy Neural Network

The adjustment of the parameters in the proposed FNN can be divided into two tasks, corresponding to the IF (antecedent) part and THEN (consequent) part of the fuzzy inference rules. A simple and intuitive method of initializing the center and width for Gaussian functions is to use normal fuzzy sets to fully cover the input space. In this paper, we initialized these singletons based on the method mentioned in[76].

A gradient-descent-based BP algorithm is employed to adjust FNN's parameters[72, 73]. The goal is to minimize the error function:

$$E = \frac{1}{2}(d - S_m)^2 \tag{5.12}$$

where S_m is the output of the FNN and d is the desired output for the input pattern. If w_{ijk} is the adjusted parameter, then the learning rule is:

$$w_{ijk}(t+1) = w_{ijk}(t) - \eta \frac{\partial E}{\partial w_{ijk}} + \alpha \Delta w_{ijk}(t) \tag{5.13}$$

and

$$\Delta w_{ijk}(t) = w_{ijk}(t) - w_{ijk}(t-1) \tag{5.14}$$

where η is the learning rate and α $(0 < \alpha < 1)$ is the momentum parameter.

In this research, the sample data for training is a McGill calibrated color Image Database[71], as mentioned in section 5.3.2. Five males who are between 20 to 30 years old are as participates in the experiments for getting teaching data. In order to obtain the teaching signals, we proceeded in the following manner. First, we got input and output data by doing lots of

experiments by using the image database mentioned above. Then got the color, intensity and orientation feature values of each image. Second, we showed the images to subjects and asked the region they intended in the experiment process. Actually, they were asked to give three regions in order. Then saliency value of pixel in users intention region was raised according to regions order. At last, by analyzing the data obtained in previous steps, we generated the teaching signals. The feature maps of image are calculated as explained above as input data. And the output data for training is the saliency value of the image calculated based on Itti's model[19] but adjusted according to the actual attention region given by user who look over the sample images.

The initial structure of the FNN uses three input nodes for x_1, x_2 and x_3, which stand for C_c, C_i and C_o, respectively. So in this case we have 3×3×3 initial rules. Suppose one epoch of learning takes 16×24 points. The supervised learning is continued for 500 epochs of training. The fuzzy sets for these linguistic term nodes are normally and uniformly initialized. We choose $\eta = 0.02$ and $\alpha = 0.85$ for supervised learning. The desired error d is got from the adjusted saliency map value calculated by Itti's method. The parameters of the initial and final membership functions are illustrated in Table 5.3. And Table 5.4 listed the weight value of the FNN after training. In the two tables A_{ijk} stands for the weight for nodes i, j, k in rule layer. Finally, the mean squared error (MSE) is 0.000497. The learning curve is illustrated in Fig.5.19. From the figure we can see that the learning speed is very fast. This is because there are only 20 groups sample data used as inputs in this time. And the number of values of each group is 16×24 points. As we all know, good production parameters can accelerate the learning speed of FNN. We also did the experiment with different settings of initial parameters by setting all to 0.5. And in this case, with the same setting of η, α and d, after 500 epochs of training, the MSE is 0.002318, which is worse than the last one.

(3)Experimental Results

After the training of FNN, We conduct several experiments to demonstrate the inference result of the proposed method and also compare with the performance of the proposed method with Itti's model[19]. As mentioned in the last section, we get the various feature

Table 5.3: Initial Parameters of the Membership Functions

Weights	Value	Weights	Value	Weights	Value
A111	0.2396	A211	0.6741	A311	0.6741
A112	0.3389	A212	0.3536	A312	0.3389
A113	0.2396	A213	0.2396	A313	0.2396
A121	0.3536	A221	0.5000	A321	0.6741
A122	0.5000	A222	0.3536	A322	0.9533
A123	0.3536	A223	0.2396	A323	0.6741
A131	0.6741	A231	0.3389	A331	0.3536
A132	0.9533	A232	0.2396	A332	0.5000
A133	0.6741	A233	0.9533	A333	0.3536

Table 5.4: Final Parameters after Learning

Weights	Value	Weights	Value	Weights	Value
A111	-0.1406	A211	0.0024	A311	0.3015
A112	0.3324	A212	0.2227	A312	0.4174
A113	0.3159	A213	0.0605	A313	0.4369
A121	0.0161	A221	0.0386	A321	0.2487
A122	0.0541	A222	0.2268	A322	0.3591
A123	0.0928	A223	0.0182	A323	-0.1952
A131	0.4472	A231	0.4159	A331	0.8196
A132	0.7299	A232	0.5501	A332	1.2614
A133	0.4659	A233	0.2168	A333	1.1184

saliency maps at first. Here, four different images are used as the input images. The input image is processed for low-level features at multiple scales, and center-surround differences are computed according to Eq.(5.4). Then, the resulting feature maps are combined into feature saliency maps according to Eq.(5.6), which is shown in Fig.5.20.

After getting the feature saliency maps, the region locations in the saliency map compete for the highest saliency value by FNN method proposed by us. After segmentation around the most salient region location, this saliency map is used for obtaining a smooth object mask at image resolution and for object-based inhibition of return. The parameters of fuzzy rules are shown in Table 5.4. The results of saliency map and attention region by both sum feature maps method and FNN method are shown in Fig.5.21. The attention regions are marked by yellow lines while red lines express the order easy to be paid attention of them.

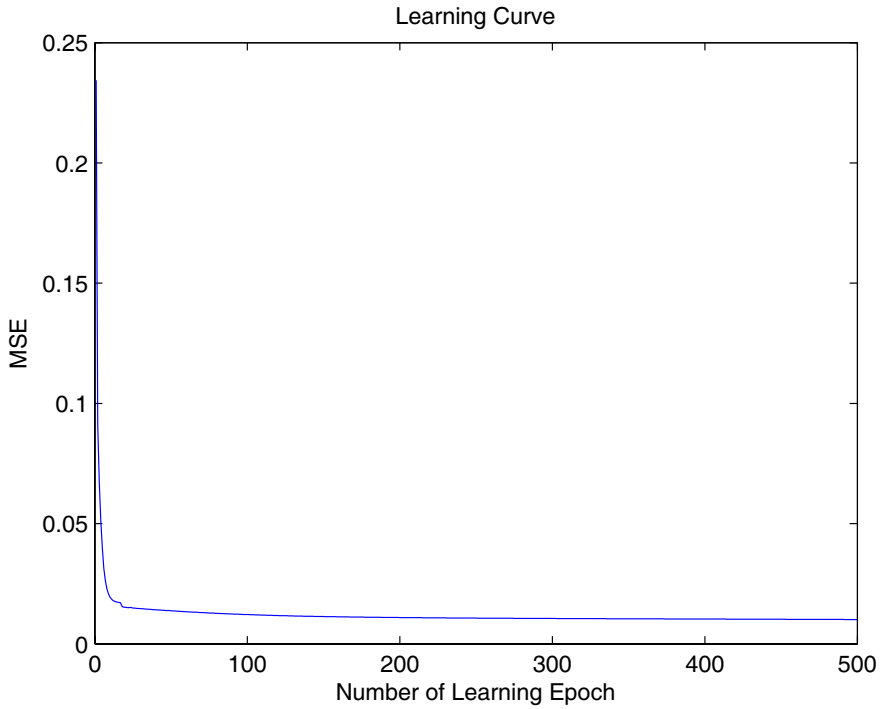


Figure 5.19: The Learning Curve of the FNN

As we can see in the figure, there are only little differences between the saliency maps of two methods. And for the first example, the approximate locations of attention regions and the orders are basically the same between two methods while the little difference is the shape and size of region. This is because the color, intensity and orientation feature are all reflect obviously in regions marked of it compared with the rest regions, which also means that the differences of importance for the three features are small. So the method we proposed has not functioned very efficiently. But from the result of the second example we can see that the attention regions of our proposed method have better result. But only from these results we cannot yet say whether our proposed method is better than Itti's[19] and fuzzy inference method or not definitely. So we conduct another experiment to verify it.

In the following experiment we asked 5 males, who are between 20 to 30 years old, to look over all of the 50 images while using the remote mode eye tracking device just as in section 5.3.2. When they looking at images, the gaze positions of them at different time

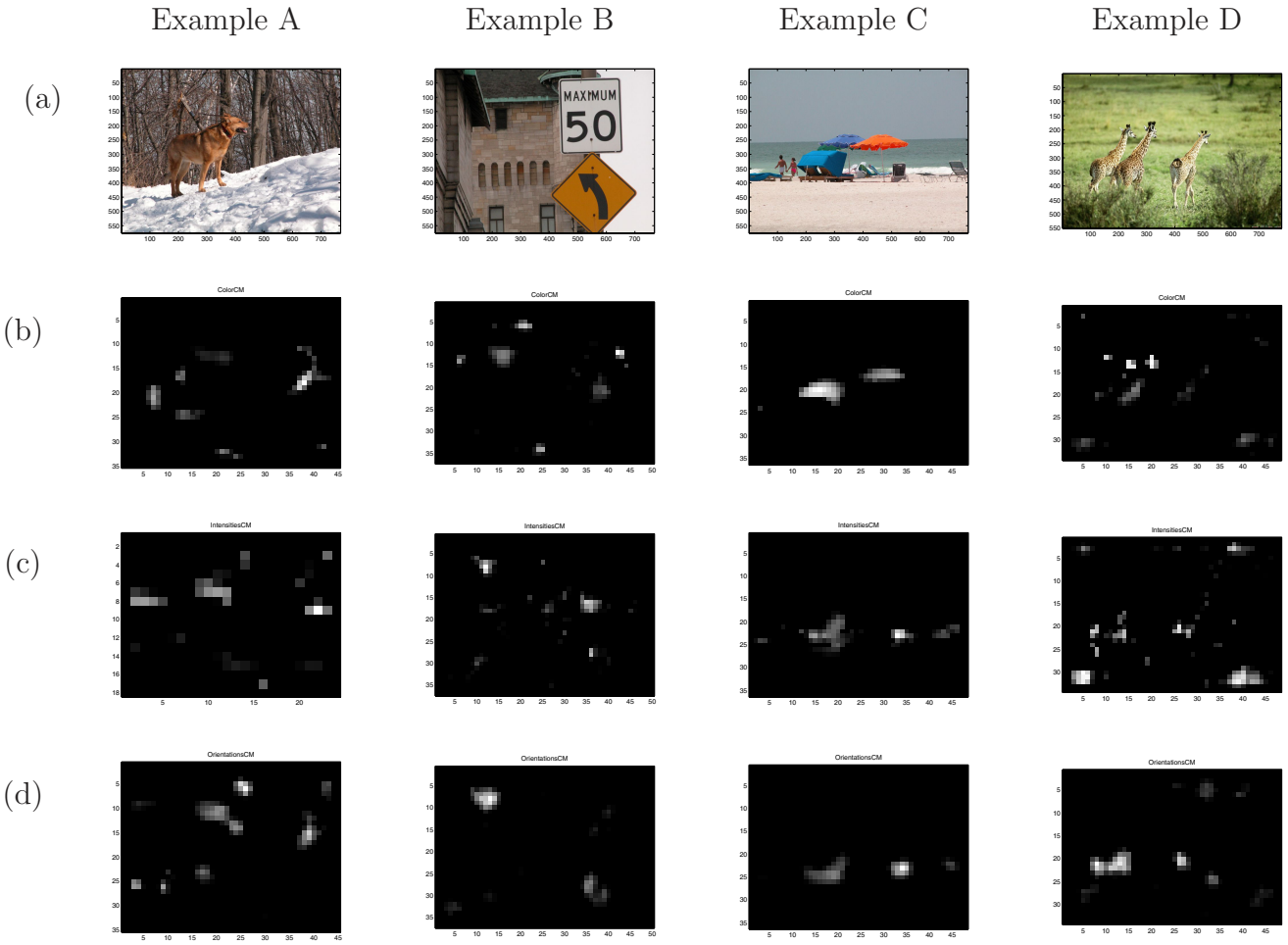


Figure 5.20: Four Examples of Feature Maps: (a)original image, (b)color feature map, (c)intensity feature map, (d)orientation feature map

are calculated and recorded for distribution analysis at last. After all the experiments they will be shown the results of attention region got by the two methods and asked to compare with the ones they actually attending and looking at in the experiment process. Finally, an evaluation of user's attitude to the results is carried on and shown in Fig.5.22. It is worth to note that the experiments method here is same with the method used in section 5.3.2. In addition, we also record the eye movements of each participant and analyze the gaze distribution. One of the results is shown in Fig.5.23. According to the evaluation results in Fig.5.22 and the gaze analysis results in Fig.5.23, we can see that the performance of our proposed method is higher than the traditional and fuzzy inference methods. This is

Figure 5.21: Four Examples of Saliency Maps and Attention Region: (a) original image, (b)attention region by sum feature maps method, (c)attention region by fuzzy inference method, (d)attention region by FNN method

also illustrated that the proposed FNN method can improve the performance of attention region prediction at some aspect.

5.4 Intention Recognition with Eye Tracking and Saliency Map

In order to combine bother subjective and non subjective for intention recognition, a visual intention region recognition system inspired on saliency map based on eye tracking is proposed. The overall procedural flow of proposed system for visual intention region

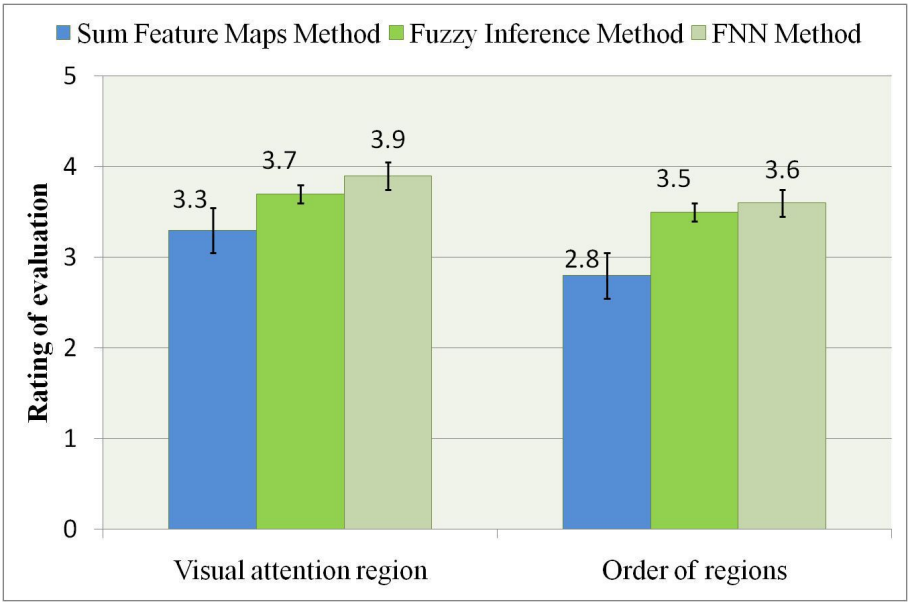


Figure 5.22: Standard Deviation for Evaluation Results of Attention Regions Obtained by Three Methods

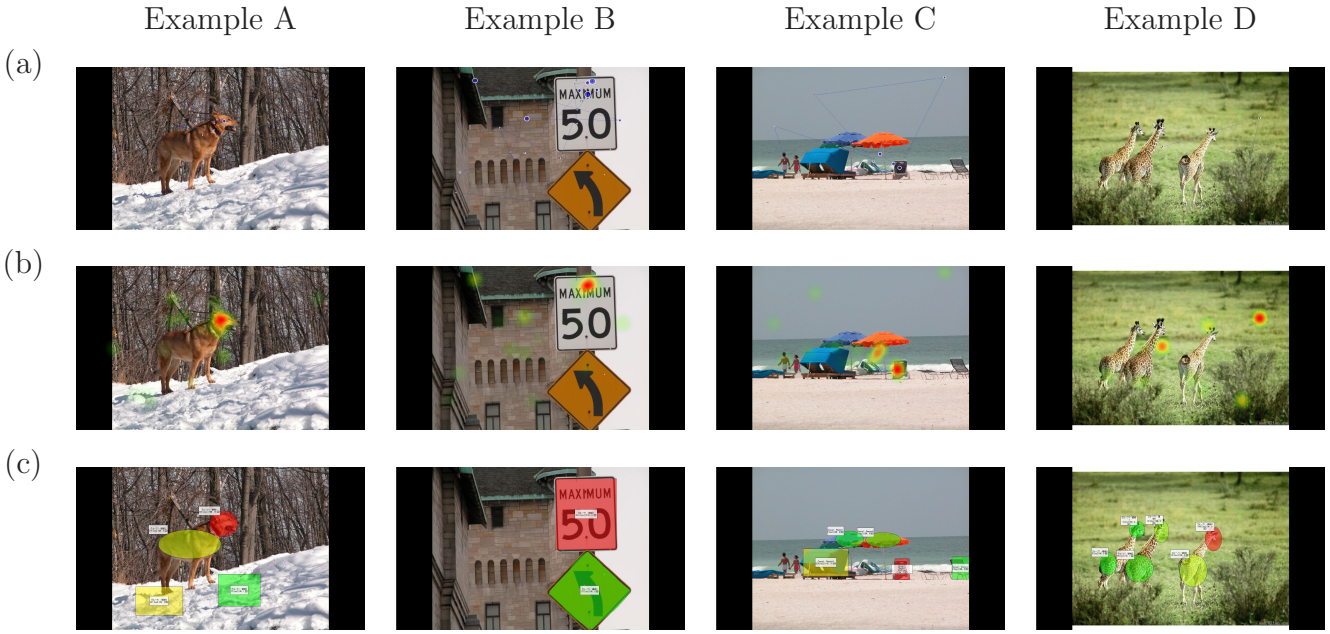


Figure 5.23: One Example of Gaze Distribution Analysis Results: (a)oder analysis, (b)stagnation map, (c)area analysis

recognition is summarized in Fig.5.24.

Firstly, the system for visual intention region recognition based on eye tracking must get real-time participant's gaze position accurately. Considering the wearable device based

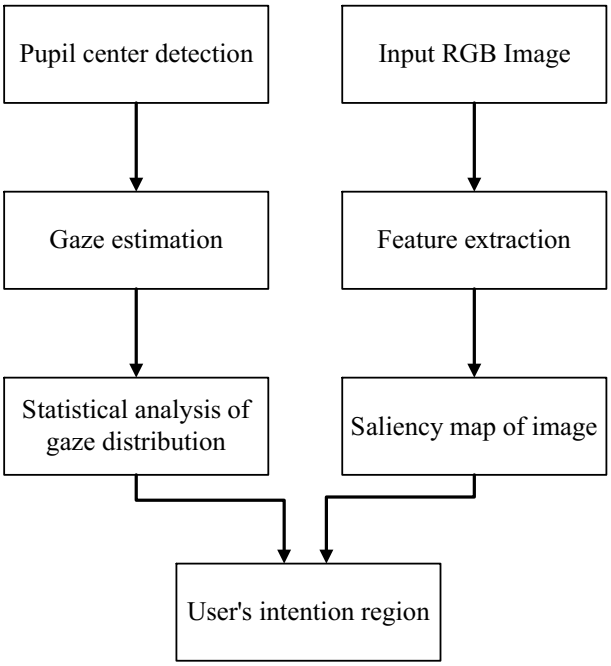


Figure 5.24: Overall Procedure for Attention Prediction with Eye Tracking and Saliency Map

eye tracking system made by ourselves has stability problem sometimes, the eye tracking system QG-PLUS Mini, described in section 4.2, is used for measuring of eye movements in this section.

Secondly, based on image processing, we infer the saliency map of the image in participant's scene by using FNN.

Thirdly, because the characteristics in the image can only reveal that the regions got through image processing and FNN mentioned above are easily to be paid attention compared with the rest ones, so the participant's attention region cannot be inferred or decided only by the characteristics in the image absolutely. Therefore, we get the participant's gaze distribution in the scene real-time by analyzing the gaze position data in a period to verify the result obtained.

Finally, the saliency map of the image is compared with the analyzed results of gaze distribution to get a more reliable result of participant's intention region. And the results of this process are used in the interactive system in following research.

5.5 Conclusions

In this chapter, firstly, we proposed an approach for intention recognition by eye tracking and object recognition. Attention can serialize learning and recognition of multiple objects in individual images. With the experiments in section 5.2.2 we show that this new mode of operation, which is impossible for the intention recognition system without prior object learning, is indeed made possible by using our saliency-based region selection algorithm.

Therefore, secondly, in section 5.3, We have limited our experiments to bottom-up attention to avoid object learning. Saliency map in this research is calculated by using color, intensity and orientation feature maps of image. However, based on the weakness of traditional saliency map calculation method that can not reflect the importance of each feature of image which is needed in generation process of saliency map, we proposed two methods based on fuzzy inference and FNN respectively. We also conducted a series of attention region prediction experiments. The prediction accuracy of our proposed methods were verified by participates' evaluation through account and analysis of eye movements by using eye tracking system. And the experimental results showed that results can be improved by using proposed methods.

Finally, participant's intention is decided not only by the characteristics of image he/she looking at, but also depends on participant's subjective factors, for example, with special interests at something or finding something. In these cases, the results obtained by proposed methods may not work well. Therefore, non subjective factors are also needed to include in the following work.

Chapter 6

Interactive System for Omnidirectional Wheelchair

6.1 Introduction

The use of power wheelchairs with high manoeuvrability and navigational intelligence is one of the great steps towards the integration of severely physically and mentally disabled people enabling them self mobility without external help. However, many of them especially those with restricted psycho-capabilities are hardly able to operate conventional wheelchairs. There are two reasons why those systems can not address these people their limited mobility and restricted functionality. Normal domestic environments are often complex structured and therefore require a high manoeuvrability. Because of their kinematic constraints conventional wheelchairs are hardly suitable to move within packed rooms. An increase of the mobility can be reached by the use of omnidirectional driving concepts.

In order to offer the people with restricted capabilities a higher degree of independence, the wheelchair has to be provided with an expanded functionality. That means beside the normal user movement by simply using a joystick several additional modes of operation which can be represented as layers within a hierarchy of functionality should be supplied to the user. For example, using basic sensor systems collision avoidance would support the user to generate safe movements. Proceeding within the hierarchy an environment guided movement can be helpful when driving along a corridor or through a door. In order to solve this issue, many methods have been proposed, such as omnidirectional vision system[77] and omnidirectional wheelchair[78].

6.2 Omnidirectional Wheelchair

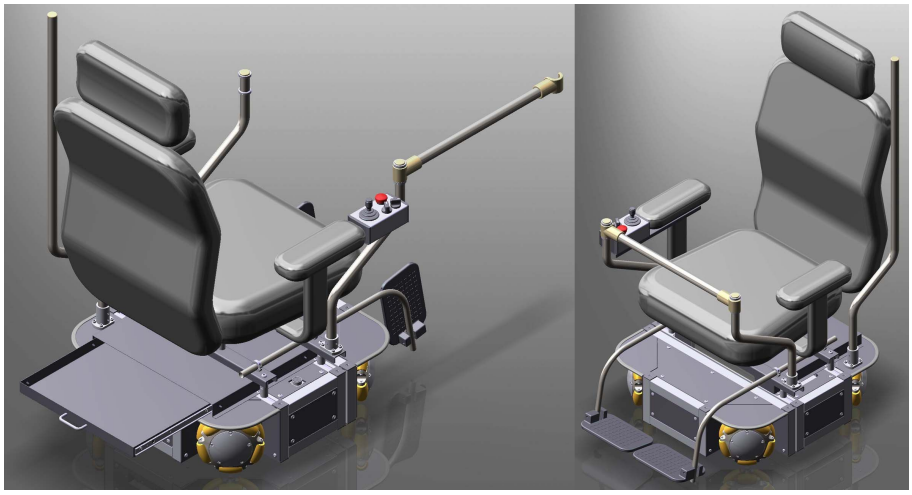


Figure 6.1: Overview of Omnidirectional Wheelchair

Recent years, omnidirectional wheelchair is researched to solve the problem mentioned in last section and has been used for various tasks. Omnidirectional wheelchair is constructed by combining traditional electric wheelchair with omni-wheels.

The omnidirectional wheelchair used in this research is as shown in Fig.6.1. The wheelchair has the ability of moving to any direction without change its own direction. The ability is achieved by using four omnidirectional wheels. And it is controlled by a digital motor controller, which is produced by maxon motor. User can also manually control it by using the joystick shown in Fig.6.1. For the convenience of programming and control, the control system of the omnidirectional wheelchair is connecting with computer by wireless. Actually, there is an adapter who provides an interface from Wifi to RS-232C. By using it, user can program and control the wheelchair in a computer, what he needs is just making sure the computer and adapter in the same local area network (LAN).

In order to control the omnidirectional wheelchair, the kinematics of the wheelchair must be known at first. The basic configuration of the four wheels of the omnidirectional wheelchair is shown in Fig.6.2. In the figure, FL, FR, BL and BR stand for the four wheels. For example, FL is the front-left wheel. The angle between each wheel's principal axis and horizontal axis is 45°. x and y are the local coordinate system of the ground plane. With the

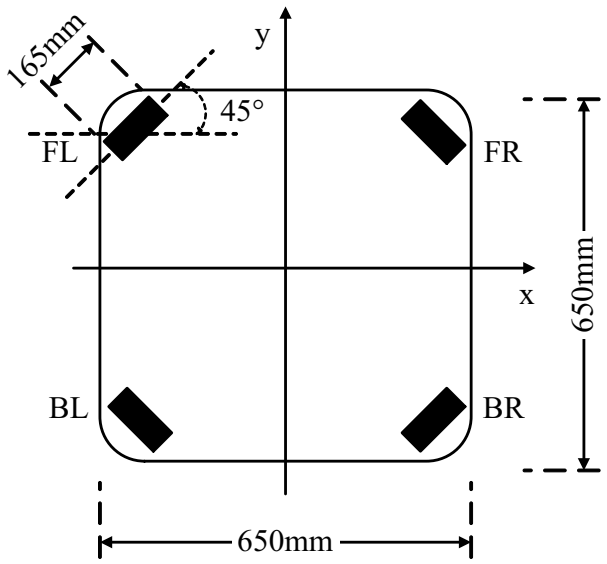


Figure 6.2: Wheels Configuration of Omnidirectional Wheelchair

configuration shown in Fig.6.2, some basic movements of the system are shown in Fig.6.3. By setting appropriate angular and speed of each wheel, the directions of movement and turning can be achieved. For example, making the omnidirectional wheelchair moving in the angle less than 45° direction can be achieved by choosing the front-left and back-right wheels available and setting the both speed of them and making sure the speed of front-left wheel is bigger than the other one.

6.3 Proposed Omnidirectional Interactive System

The structure of proposed omnidirectional interactive system based on intention recognition by using omnidirectional wheelchair is shown in Fig.6.4. In order to combine saliency map data and eye movement data, heat maps of gaze and saliency are generated after obtaining gaze position and saliency map. For saliency map, the heat map is generated by cover a mask at the saliency region. And for gaze position the heat map illustrates the region where user's gaze stays most.

In the proposed system, the eye tracking function is achieved by using QG-PLUS Mini produced by DITECT, while the saliency map is obtained by using the FNN method described in section 5.3.3.

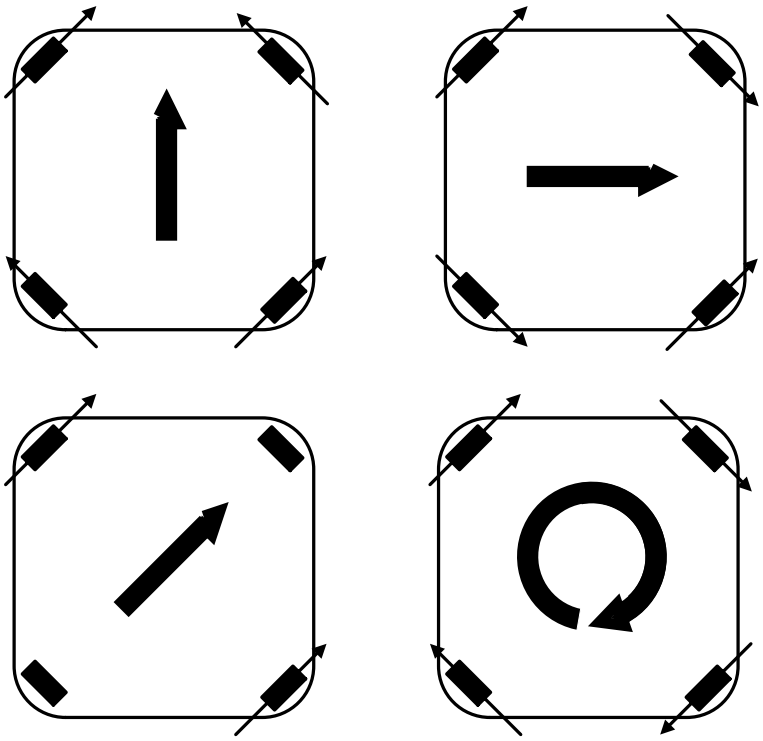


Figure 6.3: Basic Movements of the System

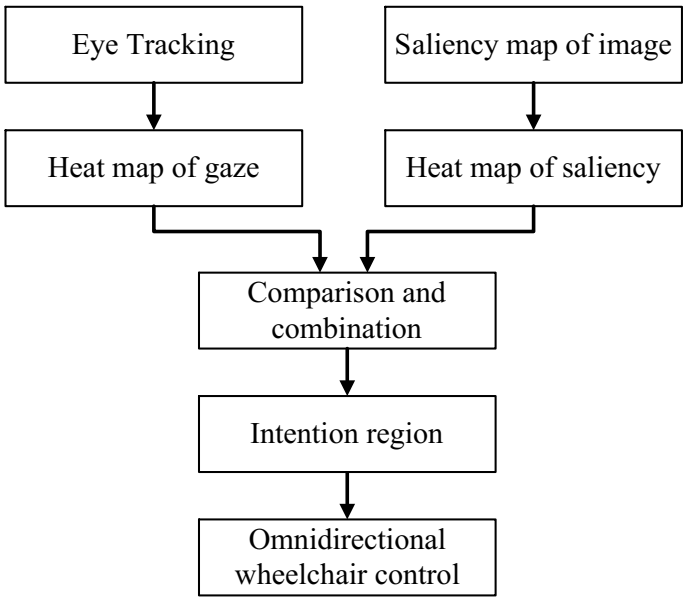


Figure 6.4: Overall Procedure for Omnidirectional Interactive System

The hardware composition is shown in Fig.6.5 and Fig.6.6. Considering the image processing speed of saliency map, for user's view scene capturing in the system we used a USB web camera with 120×160 pixels resolution. And the camera has IR light emission diode:

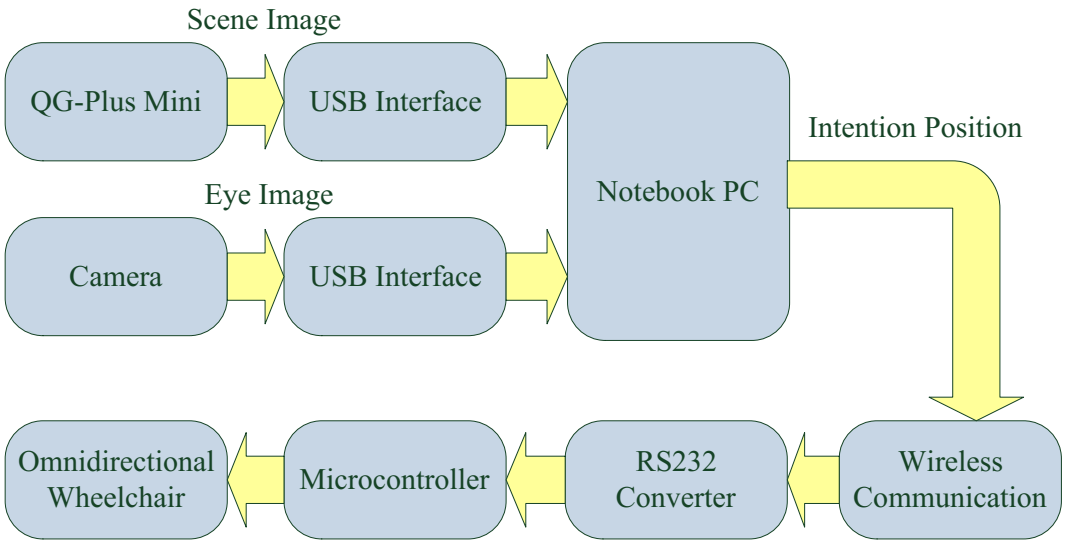


Figure 6.5: Hardware Composition of the System

LED. Therefore, it is robust against illumination changes, as shown in the position 1 in Fig.6.6. In order to shown user the scene image capturing real time, a display is set in front of the wheelchair. The eye tracking device QG-PLUS Mini is placed under the display as shown in the position 2 in Fig.6.6. And the computer used for data processing and omnidirectional wheelchair controlling is placed at the position 3 in Fig.6.6. After combining the results of saliency map and eye tracking, notebook PC sends control commands to the microcontroller of wheelchair via wireless by using RS232 converter and makes the system moving to the target area.

In the comparison and combination process, we propose the method by calculating a score of similarity between a gaze heat map and a saliency heat map. Flow of control wheelchair based on combination results of gaze map and saliency map is shown in Fig.6.7. As we can see in Fig.6.7, wheelchair is controlled according to saliency map center only when it meets the gaze map center. Otherwise, result of saliency map is considered as an untrusted result. And gaze map center will used as user's intention position. Here, the saliency heat map is achieved by giving a mask at the saliency region in saliency map. As we all know that a receiver operating characteristic or a receiver operating characteristic curve can be used for illustrating the performance of a binary classifier system as its discrimination threshold is varied[79]. It can also be used to optimise cut-off values with regard to a given

Figure 6.6: Omnidirectional Wheelchair Hardware

Table 6.1: Definition of Receiver Operating Characteristic Space

Output	Condition
True positive(TP)	P_1 is above threshold
False positive(FP)	P_2 is below threshold
True negative(TN)	P_1 is below threshold
False negative(FN)	P_2 is above threshold

prevalence in the target population and cost ratio of false-positive and false-negative results. Therefore, in our research, the space for receiver operating characteristic is defined as shown in Table.6.1.

In Table.6.1, P_1 stands for the number of pixels in region decided by gaze heat map which matches with the region of saliency heat map, while P_2 stands for the number of pixels do not match. In this case, the threshold is 100 pixels. And the true positive rate(TPR) is used as score of comparison. In a two-class prediction problem, assuming that the outputs are labeled either as positive or negative, if the output is positive and the actual value is also positive, then it is called a true positive. True positive rate, expressed as a percentage, is the probability that a test result will be positive. When the score

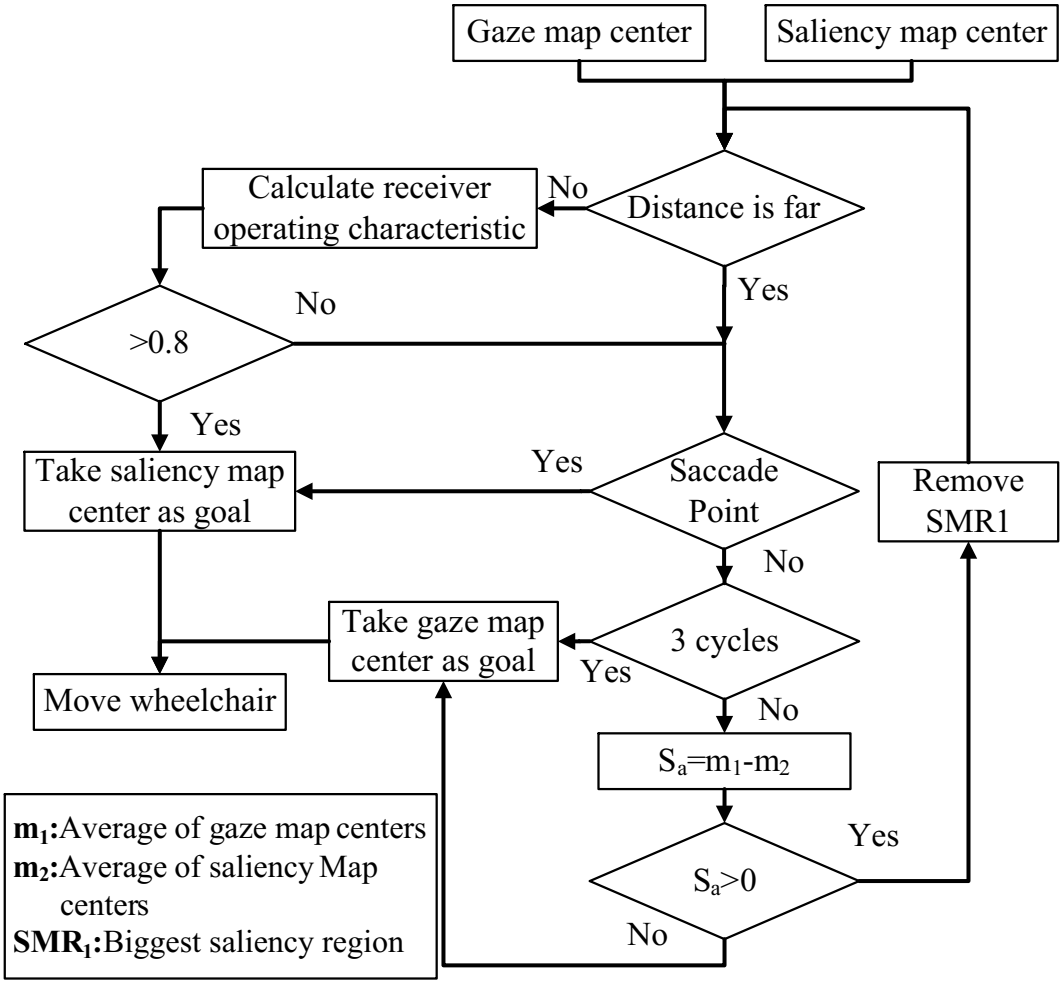


Figure 6.7: Flow of Control Wheelchair Based on Combination of Gaze Map and Saliency Map

greater than 0.8 the intention region will be confirmed or will be ignored. Then angle of the intention region center will be calculated as shown in Fig.6.8, and used to control omnidirectional wheelchair.

6.4 Experimental Result

Experiment is conducted in order to verify the proposed method. The experiment is proceeded in the following manner. User who sits on the omnidirectional wheelchair with the eye tracking device and a display in front is asked to do the eye tracking calibration process at first. After calibration, the intention region inference program starts. Following, scene image user looking at captured by the USB web camera is used as the input for saliency

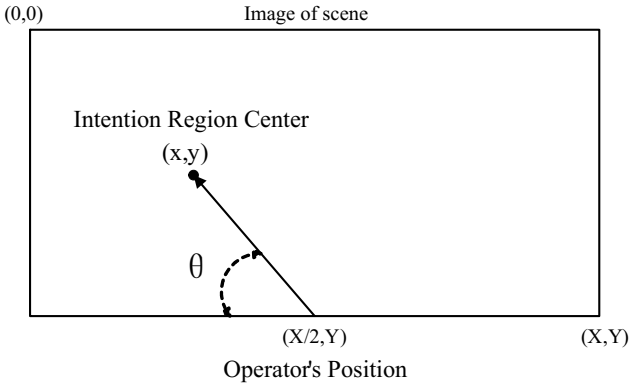


Figure 6.8: Angle Calculation of Intention Region Center

map calculation. At the same time, user's gaze position on the display is measured and recorded. Then gaze heat map and saliency heat map are generated according to saliency map and gaze distribution data. Finally, by combining gaze heat map and saliency heat map, user's intention region is decided and used to control the omnidirectional wheelchair.

Fig.6.9 shows the outlook of the proposed system, while Fig.6.11 and Fig.6.12 shows the corresponding gaze order and combined heat map for the scene image when in the status shown in Fig.6.9. In the experiment, the path of wheelchair moving according to intention recognition results is shown in Fig.6.10, corresponding to Fig.6.9. Although, for visualization purpose, Fig.6.10 shows only eight specific statements of wheelchair. From Fig.6.11 and Fig.6.12, we can see that the center of inferred saliency region is close to user's gaze position. Fig.6.10 also illustrated that wheelchair moved according to user's actual intended direction basically.

In Fig.6.11 and Fig.6.12, the left side images show the gaze order when user looking at the scene image. Note that the gaze order line is generated by analyzing gaze positions in past two seconds. The bigger the blue point shows illustrates the more time user's gaze stayed in the position. And in the right side images, the mask from blue to pink illustrating the saliency map's position while the mask from green to red standing for gaze distribution. The gaze distribution here is also generated by using gaze positions in past two seconds. The coordinate shown in heat maps is the real-time gaze coordinate of user when looking at the scene image. Ordinary, saliency map of scene image and gaze distribution show at the same time, as shown in case ⑥ in Fig.6.12. But when the gaze distribution position

Figure 6.9: Testing of Proposed System

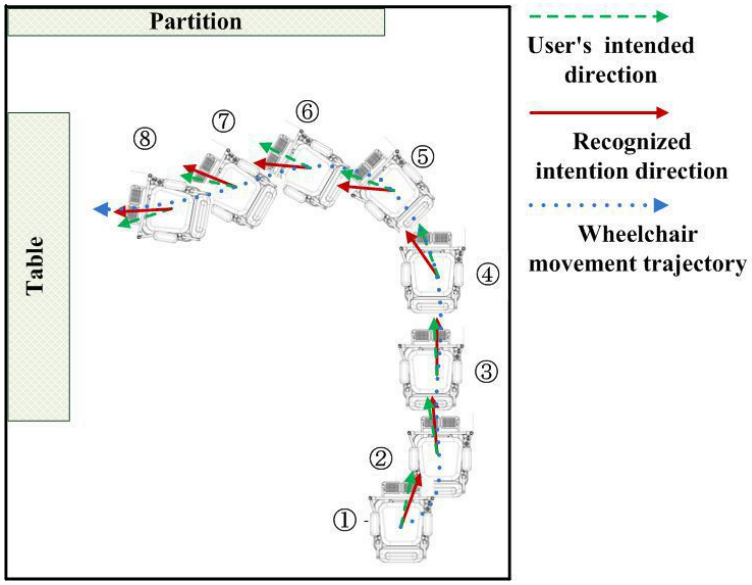


Figure 6.10: Moving Path of Wheelchair

is coincident with the saliency map position, the result will reveal the latter only. One example is shown in case ② in Fig.6.11.

Table 6.2: Gaze and Saliency Region Positions in Experiment (No.⑥ in Fig.6.12)

Gaze X	Gaze Y	Saliency Y	Saliency Y	Gaze X	Gaze Y	Saliency Y	Saliency Y
721	510	731	482	740	569	743	498
743	465	731	482	740	572	743	498
739	563	731	482	628	572	743	498
843	472	731	482	754	510	743	498
698	514	731	482	758	504	743	498
636	523	731	482	732	572	743	498
630	476	731	482	749	572	743	498
745	591	731	482	761	572	760	487
254	0	731	482	710	562	760	487
1058	65	731	482	628	570	760	487
1078	59	731	482	790	572	760	487
870	51	731	482	1001	78	760	487
749	584	743	498	765	512	760	487
732	584	743	498	759	570	760	487
723	521	743	498	769	507	760	487
598	524	743	498	473	570	760	487
694	578	743	498	781	489	760	487
710	467	743	498	710	493	760	487
786	490	743	498	745	490	760	487

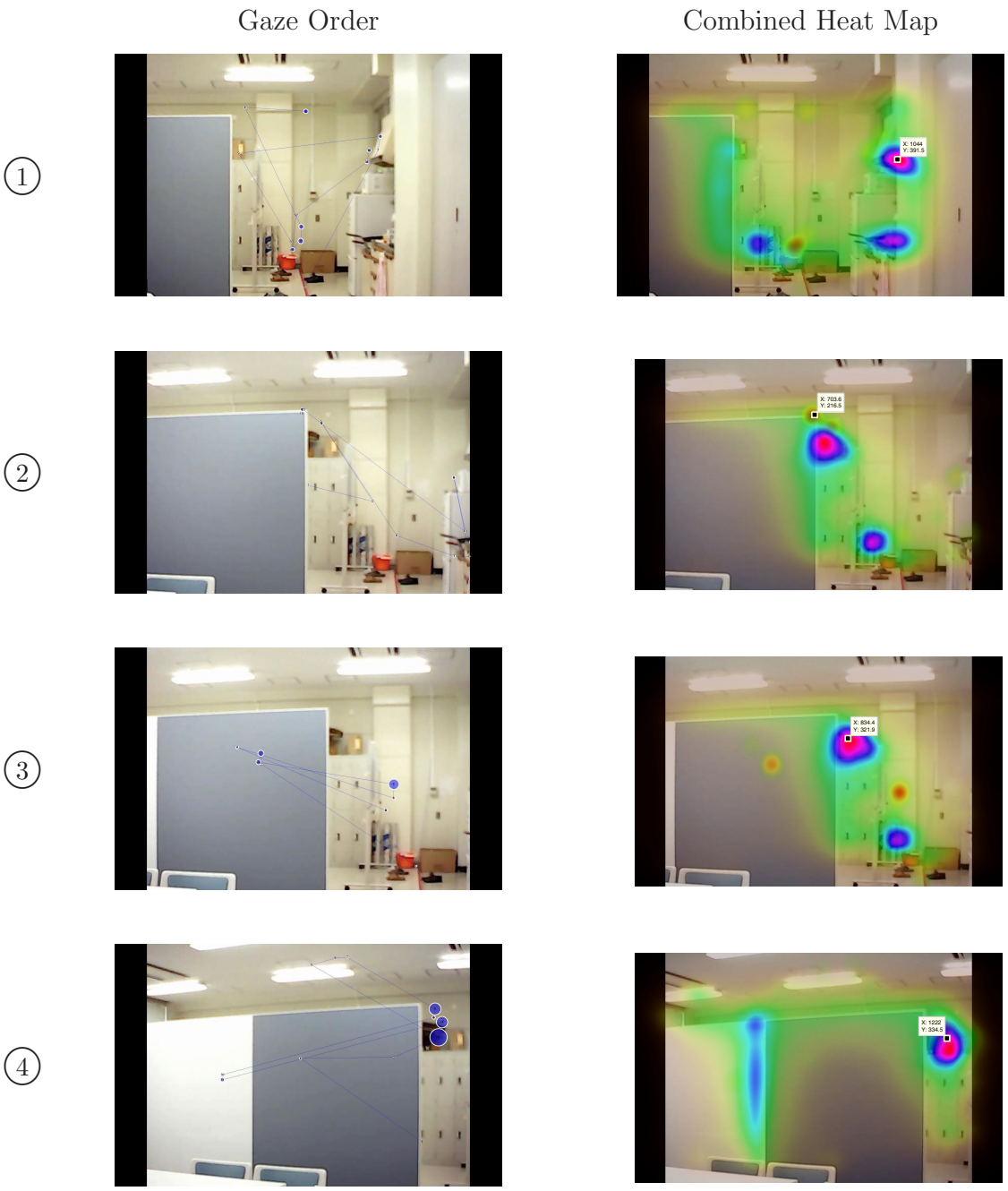


Figure 6.11: Combined Heat Map in Testing Process (1)

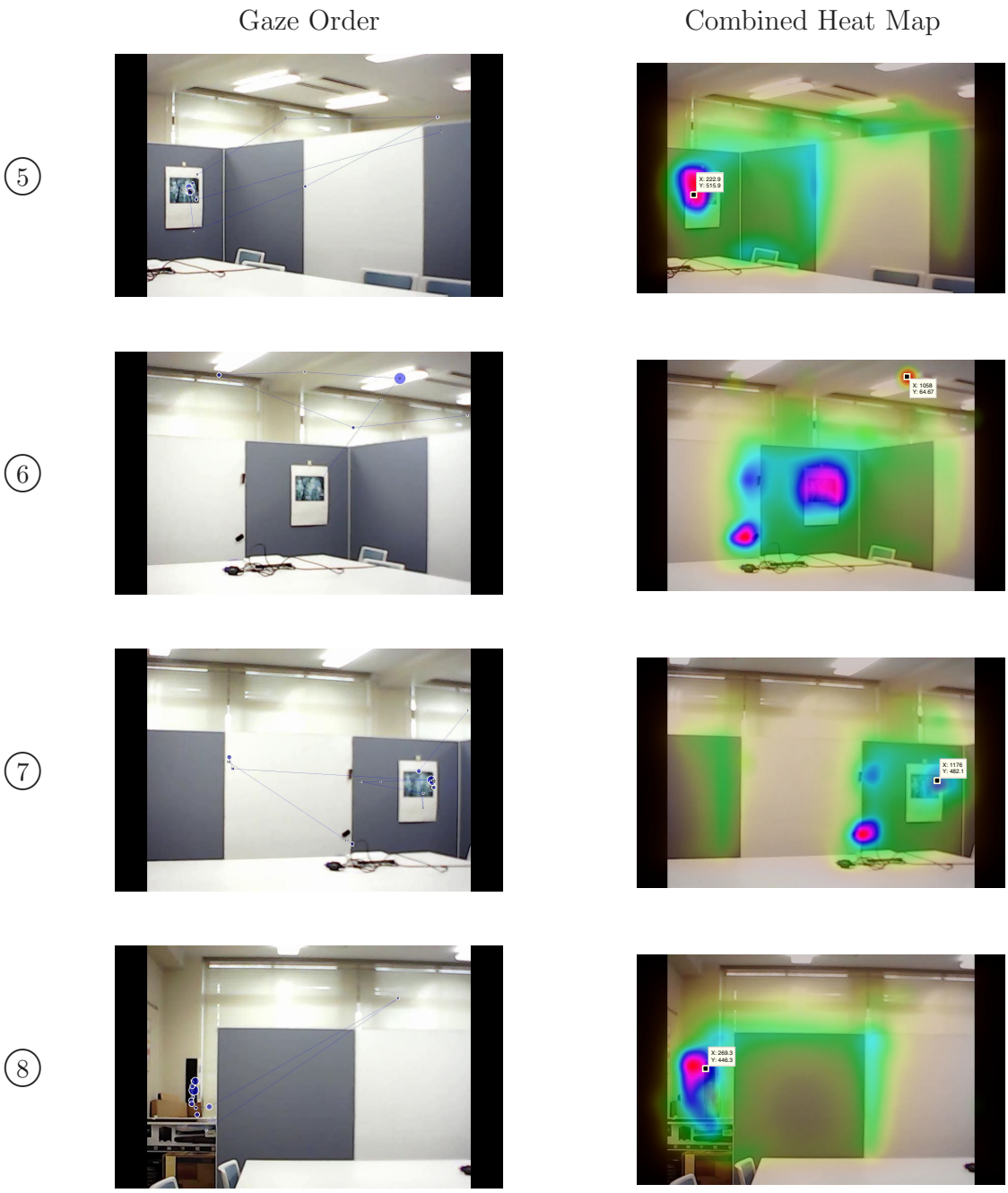


Figure 6.12: Combined Heat Map in Testing Process (2)

Table 6.3: Gaze and Saliency Region Positions in Experiment (No.⑧ in Fig.6.12)

Gaze X	Gaze Y	Saliency Y	Saliency Y	Gaze X	Gaze Y	Saliency Y	Saliency Y
273	510	240	437	110	569	273	498
281	465	240	437	244	472	273	498
269	563	240	437	245	572	273	498
210	472	240	437	267	312	273	498
897	115	240	437	304	490	273	498
310	421	240	437	309	492	273	498
467	476	240	437	271	504	273	498
301	591	240	437	276	504	270	487
230	0	240	437	239	478	270	487
238	404	240	437	300	570	270	487
236	412	240	437	313	572	270	487
236	440	240	437	241	78	270	487
238	509	273	498	257	512	270	487
312	398	273	498	268	570	270	487
321	476	273	498	279	507	270	487
322	497	273	498	276	570	270	487
322	501	273	498	274	489	270	487
310	467	273	498	310	493	270	487
274	490	273	498	312	490	270	487

6.5 Discussion

According to the results in section 6.4, we can see that in most time the gaze position is close to the center of saliency region. For example, in cases ①, ③, ④, ⑤ and ⑧ shown in Fig.6.12. Obviously, the worst result in the cases we listed here is case ⑥ shown in Fig.6.12. One of the reasons for this is that the coordinate showed here is the real time gaze position on display. The coordinate of gaze position in this case is (1058, 65).

Table.6.2 shows the gaze position data in the past two seconds. From the table we can see that the coordinate (1058, 65) is great different from other coordinates. By calculating the average of all the coordinates in the past two seconds, we know that the attention position should at the region with a coordinate about (743, 498), which matches with the center position of saliency map. Another reason is that the data processing time for saliency map is 1 second at least, which causes the display of heat map has a little delay. Based on this, we can also see that the real time gaze position coordinate may not always match with the saliency region center. To verify our analysis, gaze positions and for case ⑧ in Fig.6.12 are shown in Table.6.3. The average coordinate is (251, 439), which matches with

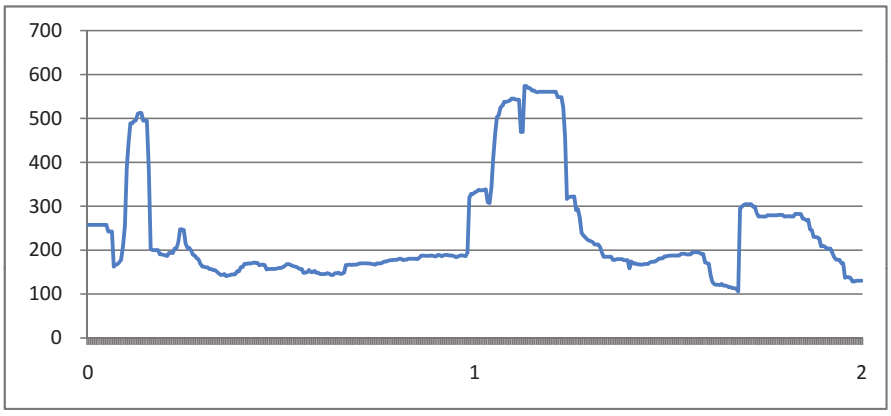


Figure 6.13: Error between Gaze Position and Saliency Region (No.⑥ in Fig.6.12)

the center of saliency region (269, 446). The error analysis results of the data in Table.6.2 and Table.6.3 are shown in Fig.6.13 and Fig.6.14, respectively. From the results we can see that in the first second user just glance at the saliency region then looked around to search other objects. And after looking around, attention has been attracted by the saliency region again. This varies with different individuals. At the first second, the distances between user's gaze position and saliency map center are about 100-200 pixels in bad results. According to actual size of the display, the distance is about 1.7-3cm. And for the good results, the distance is about 0.3-0.8cm. Considering the control of omnidirectional wheelchair is according to the angle between result position and user's position, this error is acceptable.

The analysis for Fig.6.11 and Fig.6.12 are shown in Fig.6.15.In Fig.6.15, CLL and CUL stand for confidence lower limit and confidence upper limit respectively. And confidence set in calibration process is 0.95. From the figure we can see that most of the data are in confidence interval.

6.6 Conclusions

In this chapter, an omnidirectional interactive system based on intention recognition by using omnidirectional wheelchair is proposed. The omnidirectional wheelchair is controlled by combining the results both eye tracking and saliency map. A method by using receiver operating characteristic score for comparing and combining of gaze heat map and saliency

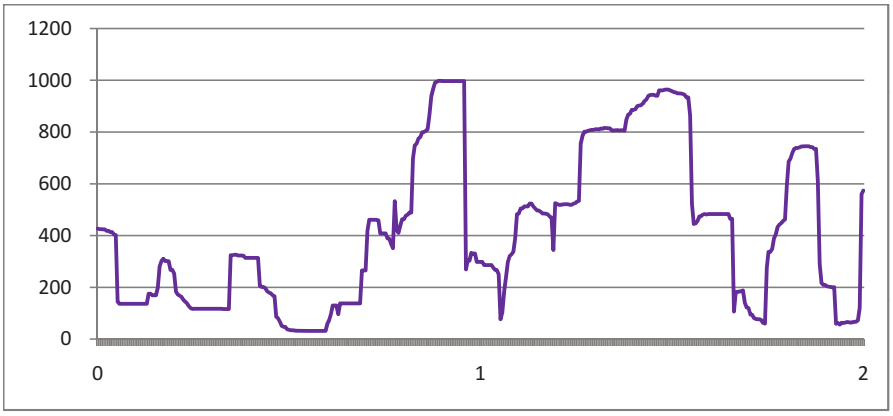


Figure 6.14: Error between Gaze Position and Saliency Region (No.⑧ in Fig.6.12)

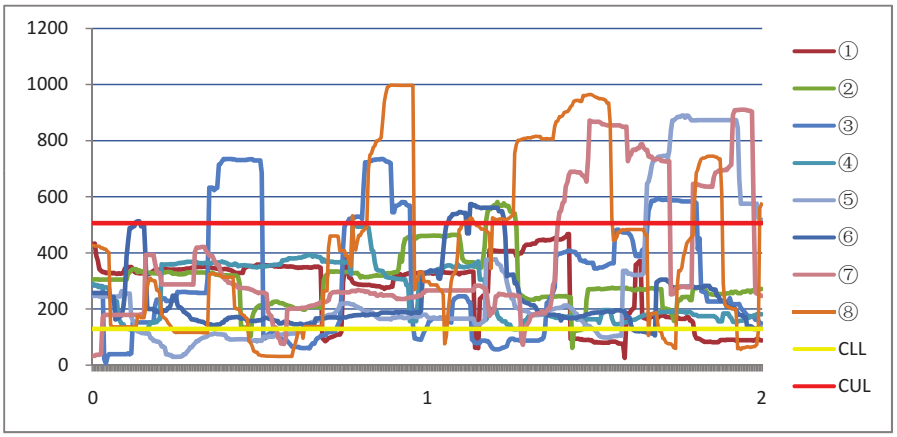


Figure 6.15: Error between Gaze Position and Saliency Region (Fig.6.11 and Fig.6.12)

heat map is also described. Different from the traditional control method of wheelchair, the method proposed can release the operator's hands and predict his/her intention at the same time. In addition, by using omnidirectional wheelchair in this system, operator is allowed to move in a confined environment with less difficulty compared with a traditional wheelchair.

Although we have set a high threshold at 0.8 for the receiver operating characteristic score, according to the testing experiment, we also found that the system may make error condition sometime. This also illuminated that for the future work, it is an important issue to improve the system by using optimal control and decision theory.

Chapter 7

Conclusions and Future Work

Since the eye tracking technology has been focused and widely researched, a lot of related and improved approaches of it were introduced and applied to many fields. Many studies on the eye tracking were mainly focused on improving the measurement accuracy of eye movement and application development in computer. Right now, the measurement accuracy is precise enough and a great diversity of corresponding applications have been put out. However, as a result of that eye movement can not really remain relatively stable like hand, gesture and other methods, it is still important to study on how to achieve an more properly method in order to make sure its wider usage. There also a lot of research worked on explore the intention recognition. Researchers made efforts to apply intention recognition in HCI, but did not get perfect effect. A typical example in this field is the research carried on by Itti based on saliency map. A lot of related researches also proved this method indeed worked.

By analyzing the factors which guide people's intention and referring to related materials, the factors that can affect one's intention can be divided into two categories: subjective factors and non-subjective ones. We argue that intention recognition just by using one of them can not obtain a result good enough. Both of the two categories of factors should be considered and combined together.

First, for application of eye tracking, we also considered there is no need for just using the coordinates of gaze as the input of an interactive system. The overall feature of it also can be considered in our research. According to this reason, in chapter 4, we described two modes of eye tracking. For the wearable device based mode, a prototype device and

corresponding software were proposed. We also try to improve the accuracy of this eye tracking system by using neural network in the calibration process, though the effect is not very obvious. Not just focus on the accuracy of the system, gaze distribution in a period and the relationship between features of eye movement and intention were also discussed. For the second mode which based on remote camera, the experiments we did are devoted to analyze gaze distribution in different ways. Related experiments for order analysis, stagnation map and area analysis when people looking at an image were conducted. These works were also prepared for the verification experiments in the following research.

Second, for the theme of visual attention with saliency map, just as mentioned in chapter 5, the method embody the importance of each feature of the image can be one of effective means to improve the results. In this chapter, two methods for calculating saliency map were proposed, the method based on fuzzy inference and the other on based on FNN. In the approach by using fuzzy inference, all of the input images had the common characteristic that color feature plays a crucial role in the saliency map. In other words, the importance of color is higher than the rest two, intensity and orientation. This was also reflected in the corresponding fuzzy rules. According to the experimental results we found that the proposed method performed better than the traditional one. But the problem was also exist which is the limitation of it. Because the input images would not always have the same characteristic that only obvious at a particular feature and the fuzzy rules also could not changed when the system was running.

To solve this issue, FNN was introduced and applied in the process of saliency map calculation. In this way, the importance of each feature can be reflected by the conceptual framework of fuzzy logic. An important point for this method is the training data for the supervised learning of FNN. In this research, a third-party color image database is imported. And the sampling of data is though involving participates in experiments.

At last, in chapter 7, we combined both the subjective and non-subjective factors together into an omnidirectional interactive system by using an omnidirectional wheelchair. The hardware specific and control method of the omnidirectional wheelchair were introduced in this chapter. Focusing on the combination method of gaze heat map and saliency heat map, a method by calculating ROC score was proposed to illustrate the performance

of two maps. Finally, the angle of intention region center was used to navigate the omni-directional wheelchair close to the object.

Research in eye tracking and intention recognition are very active. An accurate and robust measurement of eye movement can be applied into many interactive system, such as human-computer interaction, human-machine interaction and human symbiotic system. Also, with an increasing popularity of eye tracking hardware and measuring algorithms, a lot of research is being done on understanding people's intention or attention. Particularly, eye tracking is a very effective method for release our hands when interact with computer or machine. At the same time, it also the most direct reflection of our minds. Although many methods of eye tracking have been proposed, the application methods of it are fragmented and difficult to generalize. Although the method of intention prediction by eye tracking was proposed in this thesis, the features of eye movement are still used just as reference factors. In the future work, we would also focus on the prediction of gaze position to improve the proposed system. At the same time, more subjective factors, just as user's interests and so on should considered. On the other hand, an intelligent control method is also needed for the proposed system.

Acknowledgements

Foremost, I would like to express my sincere gratitude and appreciation to my supervisor, Professor Yoichiro Maeda, for his guidance, encouragement and advice he has provided throughout these past three years as his student. He has oriented and supported me with promptness and care, I have been extremely lucky to have a supervisor who cared so much about my study. His high scientific standards and hard work set an example for me. I learned a lot from him about research, how to tackle new problems and how to develop techniques to solve them.

I would like to thank Professor Kazuyuki Murase, Professor Tomohide Naniwa and Associate Professor Yasutake Takahashi for their instruction, advice and help rendered to me. I would also like to thank the professors and teachers of the graduate school of engineering, University of Fukui and Osaka Institute of Technology. I thank members of my laboratory including Kun Zhang and others, for the discussions of research and help at study and life, and for all the fun we have had in the past years.

I take this opportunity to sincerely acknowledge Ministry of Education, Culture, Sports, Science and Technology (MEXT) for providing financial assistance in the form of scholarship which buttressed me to perform my work comfortably.

Last but not least, I would like to thank my family for their emotional and moral support.

References

[1] R. Barea, L. Boquete, M. Mazo and E. L*ó*pez, "Wheelchair guidance strategies using EOG," Journal of Intelligent and Robotic Systems, Vol.34, pp.279-299, 2002.

[2] Y. Kondo, T. Miyoshi, K. Terashima and H. Kitagawa, "Navigation guidance control using haptic feedback for obstacle avoidance of omni-directional wheelchair," 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp.437-444, 2008.

[3] J. Henderson, P. Weeks Jr, and A. Hollingworth, "The effects of semantic consistency on eye movements during complex scene viewing," Journal of Experimental Psychology: Human Perception and Performance, Vol.25, No.1, pp.210-228, 1999.

[4] D. H. Yoo and M. J. Chung, "A novel non-intrusive eye gaze estimation using cross-ratio under large head motion," Computer Vision Image Understanding, Vol.98, No.1, pp.25-51, 2005.

[5] J. W. Lee, C. W. Cho, K. Y. Shin, E. C. Lee and K. R. Park, "3D gaze tracking method using Purkinje images on eye optical model and pupil," Optics and Lasers in Engineering, Vol.50, No.5, pp.736-751, 2012.

[6] P. Biswas and P. Robinson, "A brief survey on user modelling in HCI," IEEE Conference on Intelligent Human Computer Interaction, 2010.

[7] Y. Sato, M. Saito and H. Koike, "Real-time input of 3D pose and gestures of a user's hand and its applications for HCI," 2001 IEEE Virtual Reality Proceedings, pp.79-86, 2001.

[8] D. W. Hansen and Q. Ji, "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.32, No.3, pp.478-500, 2010.

[9] S. Shih, Y. Wu and J. Liu, "A calibration-free gaze tracking technique," Proceedings of the 15th international conference on pattern recognition, Vol.4, pp.4201-4205, 2000.

[10] K. Irie, B. A. Wilson, R. D. Jones, P. J. Bones and T. J. Anderson, "Laser-based eye-tracking system," Behavior Research Methods, Instruments, & Computers, Vol.34, No.4, pp.561-572, 2002.

[11] A. H. Clarke, J. Ditterich, K. Dr*ü*en, U. Sch*ö*nfeld, C. Steineke, "Using high frame rate CMOS sensors for three-dimensional eye tracking," Behavior Research Methods, Instruments, & Computers, Vol.34, No.4, pp.549-560, 2002.

[12] C. Ehmke and S. Wilson, "Identifying web usability problems from eye-tracking data," Proceedings of the 21st British HCI Group Annual Conference on People and Computers, Vol.1, pp.119-128, 2007.

[13] H. Kreinera, P. Sturtb and S. Garrod, "Processing definitional and stereotypical gender in reference resolution: Evidence from eye-movements," Journal of Memory and Language, Vol.58, No.2, pp.239-261, 2008.

[14] Y. Amit and M. Mascaro, "An integrated network for invariant visual detection and recognition," Vision Research, Vol.43, No.19, pp.2073-2088, 2003.

[15] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," Vision Research, Vol.46, pp.4333-4345, 2006.

[16] W. Einh*ä*user, C. Koch and S. Makeig, "The duration of the attentional blink in natural scenes depends on stimulus category," Vision Research, Vol.47, pp.597-607, 2007.

[17] W. Einhäuser and P. König, "Does luminance-contrast contribute to a saliency map for overt visual attention?" European Journal of Neuroscience, Vol.17, No.5, pp.1089-1097, 2003.

[18] U. Castiello, "Understanding other people's actions: intention and attention," Journal of Experimental Psychology: Human Perception and Performance, Vol.29, No.2, pp.416-430, 2003.

[19] L. Itti, "Models of bottom-up and top-down visual attention," PhD thesis, California Institute of Technology, 2000.

[20] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umilta, "Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention," Neuropsychologia, Vol.25, pp.31-40, 1987.

[21] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," Perception and Psychophysics, Vol.57, No.6, pp.787-795, 1995.

[22] M. Spain, "Modeling and predicting object attention in natural scenes," PhD thesis, California Institute of Technology, 2011.

[23] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," Vision Research, Vol.40, No.10-12, pp.1489-1506, 2000.

[24] Eye tracking and eye control for research, http://www.tobii.com/

[25] R. Rensink, J. O'Regan and J. Clark, "To see or not to see: the need for attention to perceive changes in scenes," Psychological Science, Vol.8, No.5, pp.368-373, 1997.

[26] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.20, No.11, pp.1254-1259, 1998.

[27] A. L. Yarbus, "Eye Movements During Perception of Complex Objects," Eye Movements and Vision, pp.171-211, 1967.

[28] V. R. Kenyon, "A Soft Contact Lens Search Coil for Measuring Eye Movements," Vision Research, Vol.25, No.11, pp.1629-1633, 1985.

[29] S. Shih and J. Liu, "A novel approach to 3-d gaze tracking using stereo cameras," IEEE Transactions on Systems, Man and Cybernetics, Vol.34, No.1, pp.234-245, 2004.

[30] R. Newman, Y. Matsumoto, S. Rougeaux, and A. Zelinsky, "Real-time stereo tracking for head pose and gaze estimation," Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp.122-128, 2000.

[31] D. Beymer and M. Flickner, "Eye gaze tracking using an active stereo head," IEEE Computer Vision and Pattern Recognition, pp.451-458, 2003.

[32] C. Morimoto, A. Amir, and M. Flickner, "Detecting eye position and gaze from a single camera and 2 light sources," 16th International Conference on Pattern Recognition, pp.314-317, 2002.

[33] D. Hansen and A. Pece, "Eye tracking in the wild," Computer Vision and Image Understanding, Vol.98, No.1, pp.155-181, 2005.

[34] Z. Zhu and J. Qiang, "Robust real-time eye detection and tracking under variable lighitng conditions and various face orientations," Computer Vision and Image Understanding, Vol.38, No.1, pp.124-154, 2005.

[35] C. Breazeal, "Social interactions in HRI: the robot view," IEEE Trans. Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol.34, No.2, pp.181-186, 2004.

[36] Z. Z. Bien, K. H. Park, J. W. Jung and J. H. Do, "Intention reading is essential in human-friendly interfaces for the elderly and the handicapped," IEEE Trans. Industrial Electronics, Vol.52, No.6, pp.1500-1505, 2005.

[37] S. J. Youn, K. W. Oh, "Intention recognition using graph representation," World Academy of Science, Engineering and Technology, Vol.25, pp.13-18, 2007.

[38] C. B. Kellogg and F. Zhao, "Qualitative spatial reasoning: extracting and reasoning with spatial aggregates," AI Magazine, Vol.24, No.4, pp.47-60, 2004.

[39] D. Dennett, "The Intentional Stance," MIT, 1989.

[40] C. Heinze, "Modeling Intention Recognition for Intelligent Agent Systems," PhD thesis, University of Melbourne, 2003.

[41] K. A. Tahboub, "Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition," Journal of Intelligent and Robotic Systems, Vol.45, No.1, pp.31-52, 2006.

[42] M. Wang, Y. Maeda and Y. Takahashi, "Recognition Experiment of Human Instructions Based on HMD Prototype with Eye Tracking Function," Fuzzy System Symposium 2012, pp.440-443, 2012.

[43] U. Schwarz, T. Schmuckle, "Cognitive eyes," Schweizer Archiv *für* Neurologie und Psychiatrie, Vol.153, No.4, pp.175-179, 2002.

[44] A. C. Sch*ü*tz, D. I. Braun and K. R. Gegenfurtner, "Eye movements and perception: a selective review," Journal of Vision, Vol.11, No.5, pp.1-30, 2011.

[45] D.D. Salvucci, "Inferring intent in eye-based interfaces: tracing eye movements with process models," Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp.254-261, 1999.

[46] M. Wang, Y. Maeda and Y. Takahashi, "Construction of assistant system for disabled people based on eye tracking," 31st annual conference of the robotics society of Japan, 3C2-04, 2013.

[47] A. Bulling, J. A. Ward, H. Gellersen and G. Troster, "Eye movement analysis for activity recognition using electrooculography," IEEE Tran. Pattern Analysis and Machine Intelligence, Vol.33, No.4, pp.741-753, 2011.

[48] M. Wang, Y. Maeda and Y. Takahashi, "Human Intention Recognition via Eye Tracking Based on Fuzzy Inference," Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), pp.846-851, 2012.

[49] B. Hoecks, W. Levelt, "Pupillary dilation as a measure of attention: a quantitative system analysis," Behavior Research Methods, Instruments, & Computers, Vol.25, No.1, pp.16-26, 1993.

[50] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," Pattern Analysis and Machine Intelligence, pp.970-982, 2000.

[51] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," Cognitive Psychology, Vol.12, No.1, pp. 97-136, 1980.

[52] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," Journal of Experimental Psychology: Human Perception and Performance, Vol.15, No.3, pp.419-433, 1989.

[53] C. Koch and S. Ullman, "Shifts in selection in visual attention: Towards the underlying neural circuitry," Human Neurobiology, Vol.4, No.4, pp.219-227, 1985.

[54] D. Liu and T. Chen, "DISCOV: A framework for discovering objects in video," IEEE Trans. Multimedia, Vol.10, No.2, pp.200-208, 2008.

[55] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," Journal of Vision, Vol.9, No.3, pp.1-24, 2009.

[56] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," Journal of Vision, Vol.9, No.12, pp.1-27, 2009.

[57] A. D. Hwang, E. C. Higgins, and M. Pomplun, "A model of top-down attentional control during visual search in complex scenes," Journal of Vision, Vol.9, No.5, pp.1-18, 2009.

[58] Z. Zhu, Q. Ji, K. Fujimura and K. Lee, "Combining Kalman filtering and mean shift for real time eye tracking under active IR illumination," Proceedings of the 16th Pattern Recognition International Conference, Vol.4, pp.318-321, 2002.

[59] C. H. Morimoto, D. Koons, A. Amir, M. Flickner, "Pupil detection and tracking using multiple light sources," Image and Vision Computing, Vol.18, No.4, pp.331-335, 2000.

[60] K. M. Lam, H. Yan, "Locating and extracting eye in human face images," Pattern Recognition, Vol.29, No.5, pp.771-779, 1996.

[61] T. Rajpathaka, R. Kumarb and E. Schwartzb, "Eye detection using morphological and color image processing," Proceedings of the Florida Conference on Recent Advances in Robotics, pp.1-6, 2009.

[62] DITECT, http://www.ditect.co.jp/

[63] K. A. Tahboub, "Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition," Journal of Intelligent and Robotic Systems, Vol.45, pp.31-52, 2006.

[64] A. Bykat, "Convex hull of a finite set of points in two dimensions," Info. Proc. Letters, Vol.7, pp.296-298, 1978.

[65] I. Mitsugami, N. Ukita and M. Kidode, "Robot Navigation by Eye Pointing," Proc. Entertainment Computing, pp.256-267, 2005.

[66] Documentation of Aldebaran Nao from Aldebaran Robotics -Documentation Version, 1.14.3.

[67] W. O. Lee, J. W. Lee, K. R. Park, E. C. Lee and M. Whang, "Object recognition and selection method by gaze tracking and SURF algorithm," 2011 International Conference on Multimedia and Signal Processing, pp.261-265, 2011.

[68] J. W. Harris and H. Stocker, Handbook of Mathematics and Computational Science, Springer-Verlag New York, 1998.

[69] D. Walther, U. Rutishauser, C. Koch and P. Perona, "On the usefulness of attention for object recognition," Workshop on Attention and Perfromance in Computational Vision, pp.96-103, 2004.

[70] M. Wang, Y. Maeda and Y. Takahashi, "Visual Attention Region Prediction Based on Eye Tracking Using Fuzzy Inference," Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.18, No.4, 2014.

[71] A. Olmos, F. A. A. Kingdom, "A biologically inspired algorithm for the recovery of shading and reflectance images," Perception, Vol.33, No.12, pp.1463-1473, 2004.

[72] G. B. Huang, Q. Y. Zhu and C. K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference, Vol.2, pp.985-990, 2004.

[73] C. Chao, C. Teng, "Implementation of a fuzzy inference system using a normalized fuzzy neural network," Fuzzy Sets and Systems, Vol.75, No.1, pp.17-31, 1995.

[74] C. M. Lin, C. F. Hsu, "Supervisory recurrent fuzzy neural network control of wing rock for slender delta wings," IEEE Transactions on Fuzzy Systems, Vol.12, No.5, pp.733-742, 2004.

[75] M. Wang, Y. Maeda and Y. Takahashi, "Saliency Map for Visual Attention Region Prediction Based on Fuzzy Neural Network," WCCI 2014 IEEE World Congress on Computational Intelligence, F-14407, 2014.

[76] M. Wang, Y. Maeda and Y. Takahashi, "A Fuzzy Inference Method Based on Saliency Map for Visual Attention Region Prediction," Fuzzy System Symposium 2013, pp.495-500, 2013.

[77] Y.Maeda and W.Shimizuhira, "Multi-Layered Fuzzy Behavior Control for Autonomous Mobile Robot with Multiple Omnidirectional Vision System: MOVIS," Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.11, No.1, pp.21-27, 2007.

[78] J. Urbano, K. Terashim, T. Miyosh and H. Kitagawa, "Collision avoidance in an omnidirectional wheelchair by using haptic feedback," 4th WSEAS International Conference on Signal Processing, Robotics and Automation, No.18, 2005.

[79] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, Vol.27, No.8, pp.861-874, 2006.